(continued from issue 4)

Computer languages

The main difference between a computer and, say, an ordinary calculator is that a computer can be **programmed** to automatically carry out operations on data which is fed into it. In a pocket calculator only very small programs can be stored for automatic operation on data. These operations are generally written by the programmer in what is known as a high level language. This is a language that is understandable by people as well as computers and so makes programming relatively easy.

When a program written in a high level language is run in a computer, a part of the computer's operating system – known as the compiler – translates it into sets of binary configurations that the computer can understand and carry out as operations. These instructions constitute its **machine language.** Each machine language instruction represents an elementary computer operation.

Every high level language instruction usually corresponds to several machine language instructions. Each computer has a particular set of instructions which makes up its machine language and they are usually classified into the groups shown in *table 5*. These instructions cover the computer's arithmetical, logical, control transfer, data movement and machine operation commands.

The big general purpose computers have these and many other instructions in their machine language set. The most economical CPUs, such as used in mini and microcomputers, generally have a smaller, less sophisticated set. So to perform the same processing functions as the bigger computers, they will have to use more machine instructions, and this takes more time to carry out.

Each instruction is identified by an **operation code** (op code). This tells the CPU exactly which operation it is to perform.

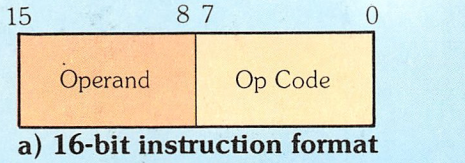
The size of machine instructions (in terms of the number of bits used to represent them) varies depending on the type of computer and the addressing techniques it uses. **Addressing** means the process of locating the information (**operand**) necessary for a particular operation.

Instructions can take up one or more memory words, and the length of the words themselves can vary from computer to computer. *Figure 2* shows two different instruction formats. In the first example 16 bits are used to specify an 8 bit operating code and an 8 bit operand. The operand is located in the instruction. When operands are part of the instruction the technique is called **immediate addressing.** That is, the operand is immediately available once the instruction has been fetched. The op code indicates what is to be done to the operand. For example, it may be used to change the program counter or perhaps to add to the contents of another register.

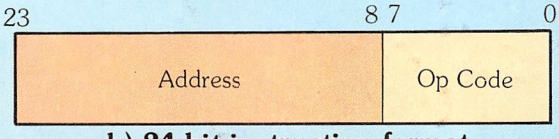
With an 8 bit word, the number of different bit combinations available is 256

2. How instructions are contained in memory cells. The 16-bit format, with the operand located in the instruction, is known as immediate addressing. The 24-bit format, which specifies the memory address of the operand, is referred to as direct addressing.

2



a) 16-bit instruction format



b) 24-bit instruction format

Table 5
Subdivision of machine instructions

Arithmetic	(Add, subtract, multiply, divide, and negation)
Logical	(AND, OR, NOT, and EXCLUSIVE OR)
Transfer of Control	(Unconditional branching, conditional branching, loops and subroutines)
Data Movement, Input and Output, Data Assignment	(Move, load, and store)

Some typical machine instructions are:

LOAD	Register from memory
STORE	Register to memory
MOVE	Register to register
Add, Subtract, Multiply, Divide	Arithmetic operations on register contents with other register or memory contents
AND, OR, NOT	Logical operations on registers or memory contents

Table 6
6-bit alphanumeric codes

Value of 4 least significant bits	Value of 2 most significant bits											
	Code IBM 1400				Code CDC 3000				Code H 6000			
	00	01	10	11	00	01	10	11	00	01	10	11
0 0 0 0	b		—	+	0	+	—	b	0	b	↑	+
0 0 0 1	1	/	J	A	1	A	J	/	1	A	J	/
0 0 1 0	2	S	K	B	2	B	K	S	2	B	K	S
0 0 1 1	3	T	L	C	3	C	L	T	3	C	L	T
0 1 0 0	4	U	M	D	4	D	M	U	4	D	M	U
0 1 0 1	5	V	N	E	5	E	N	V	5	E	N	V
0 1 1 0	6	W	O	F	6	F	O	W	6	F	O	W
0 1 1 1	7	X	P	G	7	G	P	X	7	G	P	X
1 0 0 0	8	Y	Q	H	8	H	Q	Y	8	H	Q	Y
1 0 0 1	9	Z	R	I	9	I	R	Z	9	I	R	Z
1 0 1 0	0						°	&	[	&	—	←
1 0 1 1	#	,	$	.	=	.	$	,	#	.	$	,
1 1 0 0	@	%	*	◊	"	)	*	(	@	]	*	%
1 1 0 1					:	'	#		:	(	)	=
1 1 1 0					;	@	%		>	<	;	"
1 1 1 1					?	!			?		'	!

(2[8]). The operand in this case could be any value between 0 and 255, or a number between -128 and +127, (as one bit is usually used to indicate the sign of a number).

In the second example of an instruction format in *figure 2*, 24 bits are used. Here the first 8 bits are used to specify the op code. The next 16 bits specify the memory address of the operand. This is known as **direct addressing.** The operand can be found in one of 65,536 memory locations (65,536 = 24K = 2[16]).

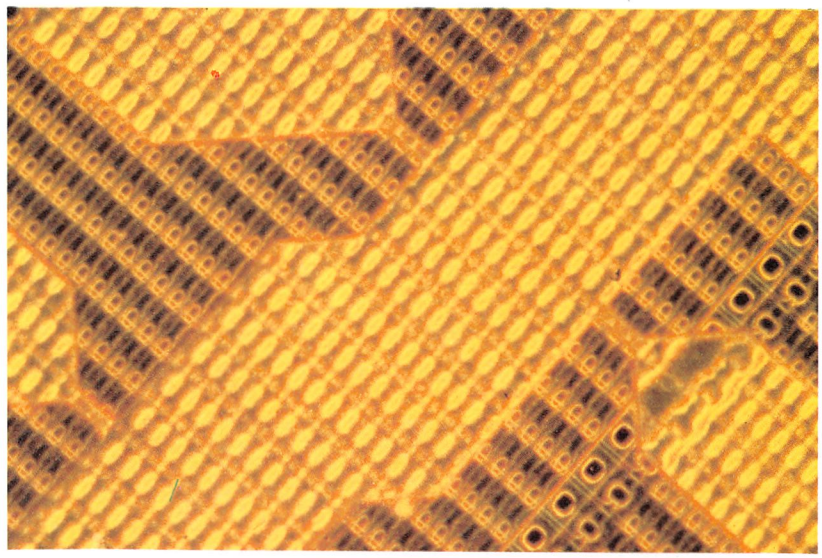
How information is represented in binary codes

Computers can represent any type of information in binary codes. We will see how the same bit configuration can be interpreted in different ways depending on the nature of the different codes used. All the alphanumerical and graphics characters that a computer can read in or print out are known as its **external alphabet.** It usually consists of:
1) the 26 letters of the alphabet, lower case, upper case or both
2) the ten decimal digits 0 to 9
3) various other characters such as punctuation marks, operational symbols, etc.

Different codes exist for the representation of these characters. A widespread code for numerical characters is the BCD (binary coded decimal) code. It uses 4 bits to represent the decimal digits from 0 to 9 and portrays a number by decoding each digit separately.
For example, 1984 in BDC is:

Decimal	1	9	8	4
BCD	0001	1001	1000	0100

Obviously, encoding a number in BCD requires more bits than it would in binary: moreover 16 characters can be represented with 4 bits and not just the 10 of the decimal system. From this point of



Part of a silicon chip used in IBM terminals (greatly magnified). (72,000 pieces of information are contained in just ¼ inch.)

view BCD may not seem to be the best use of resources, but it is used by printing devices because the decoding circuits can be made very simple.

For example let's look at the addition of 7 and 6 in BCD:

```
0111        7
0110+       6+
----        --
1101        13
```

However 1101 is not a BCD character. 13 coded in BCD should be 0001 0011.

Computers which use BCD have an instruction called 'decimal adjustment' in their machine language. This makes sure that each 4 bit word cannot indicate a number larger than 9 in BCD.

As we have mentioned, computers have to have some way of representing characters other than numbers. **Alphanumeric** codes allow all the external alphabet of the computer to be portrayed in binary form. 4 bit words are not suitable for this purpose as they only allow 16 (2^4) different combinations of bits to exist. This would be useless if we wanted to represent the letters of the alphabet, so codes of 6, 7 and 8 bits are used.

Table 6 shows some of the more commonly used alphanumeric codes. With 6 bits, 64 (2^6) symbols can be encoded. These are generally the 26 capital letters of the alphabet, the 10 decimal digits and 28 various characters. The 7 bit ISO code (International Standards Organisation) allow 128 (2^7) characters to be encoded (*See table 7*).

Two very widely used 8 bit codes are EBCDIC (Extended Binary Code Decimal Interchange Code) and ASCII (American Standard Code for Information Interchange). These can allow the representation of up to 256 (2^8) characters, and are shown in *tables 8* and *9*.

Table 7
7 bit ISO code

Value of 4 least significant bits	Value of 3 MSB 000	001	010	011	100	101	110	111
0 0 0 0			b	0	∝	P	'	p
0 0 0 1			!	1	A	Q	a	q
0 0 1 0			"	2	B	R	b	r
0 0 1 1			£	3	C	S	c	s
0 1 0 0			$	4	D	T	d	t
0 1 0 1			%	5	E	U	e	u
0 1 1 0			&	6	F	V	f	v
0 1 1 1			'	7	G	W	g	w
1 0 0 0			(	8	H	X	h	x
1 0 0 1			)	9	I	Y	i	y
1 0 1 0			*	:	J	Z	j	z
1 0 1 1			+	;	K	[	k	
1 1 0 0			,	<	L	\	l	
1 1 0 1			—	=	M	]	m	
1 1 1 0			.	>	N	↑	n	
1 1 1 1			/	?	O	—	o	

Table 8
ASCII code

Dec	Hex	Chr	Dec	Hex	Chr	Dec	Hex	Chr	Dec	Hex	Chr	Dec	Hex	Chr	Dec	Hex	Chr	Dec	Hex	Chr
0	0	NUL	19	13	DC3	38	26	&	57	39	9	76	4C	L	95	5F	—	114	72	r
1	1	SOH	20	14	DC4	39	27	'	58	3A	:	77	4D	M	96	60	'	115	73	s
2	2	STX	21	15	NAK	40	28	(	59	3B	;	78	4E	N	97	61	a	116	74	t
3	3	ETX	22	16	SYN	41	29	)	60	3C	<	79	4F	O	98	62	b	117	75	u
4	4	EOT	23	17	ETB	42	2A	*	61	3D	=	80	50	P	99	63	c	118	76	v
5	5	ENQ	24	18	CAN	43	2B	+	62	3E	>	81	51	Q	100	64	d	119	77	w
6	6	ACK	25	19	EM	44	2C	,	63	3F	?	82	52	R	101	65	e	120	78	x
7	7	BEL	26	1A	SUB	45	2D	—	64	40	@	83	53	S	102	66	f	121	79	y
8	8	BS	27	1B	ESC	46	2E	.	65	41	A	84	54	T	103	67	g	122	7A	z
9	9	HT	28	1C	FS	47	2F	/	66	42	B	85	55	U	104	68	h	123	7B	{
10	A	LF	29	1D	GS	48	30	0	67	43	C	86	56	V	105	69	i	124	7C	:
11	B	VT	30	1E	RS	49	31	1	68	44	D	87	57	W	106	6A	j	125	7D	}
12	C	FF	31	1F	US	50	32	2	69	45	E	88	58	X	107	6B	k	126	7E	~
13	D	CR	32	20	SP	51	33	3	70	46	F	89	59	Y	108	6C	l	127	7F	DEL
14	E	SO	33	21	!	52	34	4	71	47	G	90	5A	Z	109	6D	m			
15	F	S1	34	22	"	53	35	5	72	48	H	91	5B	[	110	6E	n			
16	10	DLE	35	23	#	54	36	6	73	49	I	92	5C	\	111	6F	o			
17	11	DC1	36	24	$	55	37	7	74	4A	J	93	5D	]	112	70	p			
18	12	DC2	37	25	%	56	38	8	75	4B	K	94	5E	^	113	71	q			

Note: Dec = Decimal, Hex = Hexadecimal, Chr = Character (The first 32 codes are for control functions)

How instructions control a computer
Let's look in detail at the way in which a computer obeys the instructions in a program. For example, take an instruction to load an item of data from the memory to an internal register of the CPU. The machine code for this instruction could be the hexadecimal number 3E 0F, which corresponds to the binary number 0011 1110 0000 1111. The instruction is situated in memory, starting from the location addressed 01FE. This is shown on the right of *figure 3*.

The number 3E represents the operation code of the instruction and it is found in the memory at the address 01FE. The number 0F is the operand and it is found in the memory at the address 01FF. (For convenience, address numbers are expressed in the hexadecimal system.)

When a program is run, the address of the first instruction is loaded into the **program counter** (PC). This is a register in the CPU which stores the address of the next instruction to be performed.

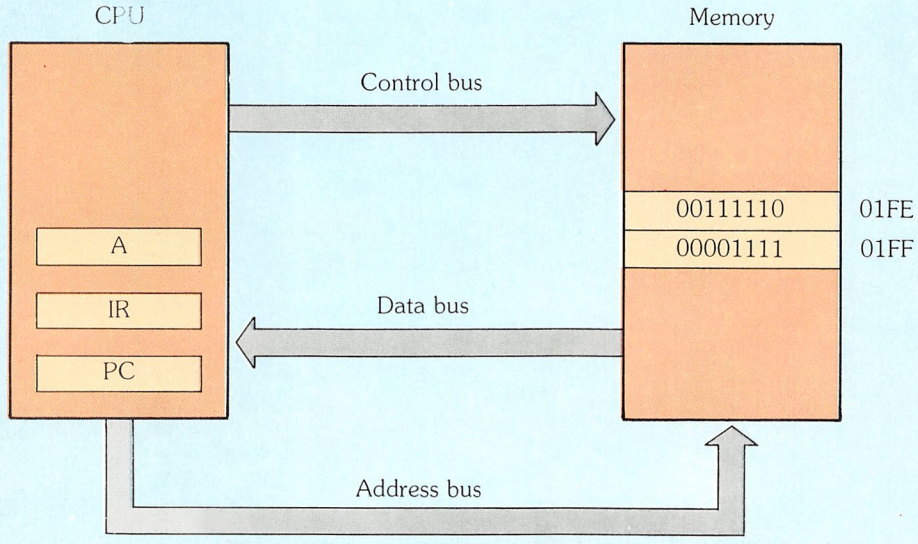
Imagine that our stored instruction is the beginning of a program. Its memory address (01FE) will be found in the program counter.

Now look at what happens step by step:
1) the content of the program counter, i.e.

Table 9
8 bit EBCDIC code

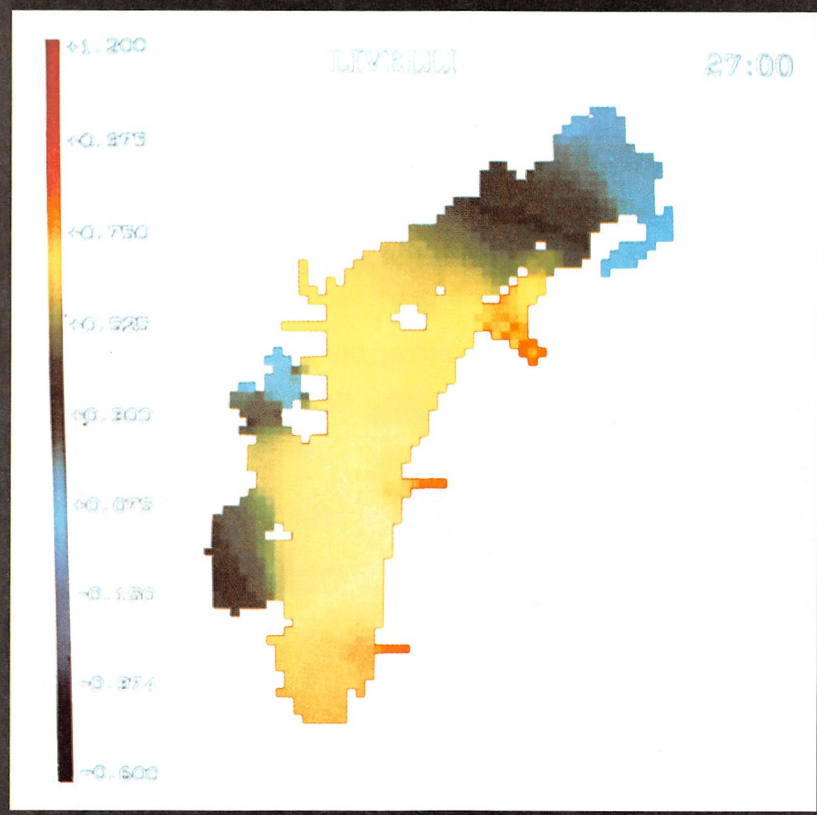
s	code	s	code	s	code	s	code
b	0100 0000	a	1000 0001	A	1100 0001	∅	1111 0000
c	0100 1010	b	1000 0010	B	1100 0010	1	1111 0001
°	0100 1011	c	1000 0011	C	1100 0011	2	1111 0010
<	0100 1100	d	1000 0100	D	1100 0100	3	1111 0011
(	0100 1101	e	1000 0101	E	1100 0101	4	1111 0100
+	0100 1110	f	1000 0110	F	1100 0110	5	1111 0101
	0100 1111	g	1000 0111	G	1100 0111	6	1111 0110
		h	1000 1000	H	1100 1000	7	1111 0111
&	0101 0000	i	1000 1001	I	1100 1001	8	1111 1000
!	0101 1010					9	1111 1001
$	0101 1011	j	1001 0001	J	1101 0001		
*	0101 1100	k	1001 0010	K	1101 0010		
)	0101 1101	l	1001 0011	L	1101 0011		
;	0101 1110	m	1001 0100	M	1101 0100		
¬	0101 1111	n	1001 0101	N	1101 0101		
		o	1001 0110	O	1101 0110		
—	0110 0000	p	1001 0111	P	1101 0111		
/	0110 0001	q	1001 1000	Q	1101 1000		
'	0110 1011	r	1001 1001	R	1101 1001		
%	0110 1100						
—	0110 1101	s	1010 0010	S	1110 0010		
>	0110 1110	t	1010 0011	T	1110 0011		
?	0110 1111	u	1010 0100	U	1110 0100		
		v	1010 0101	V	1110 0101		
:	0111 1010	w	1010 0110	W	1110 0110		
#	0111 1011	x	1010 0111	X	1110 0111		
@	0111 1100	y	1010 1000	Y	1110 1000		
'	0111 1101	z	1010 1001	Z	1110 1001		
=	0111 1110						
"	0111 1111						

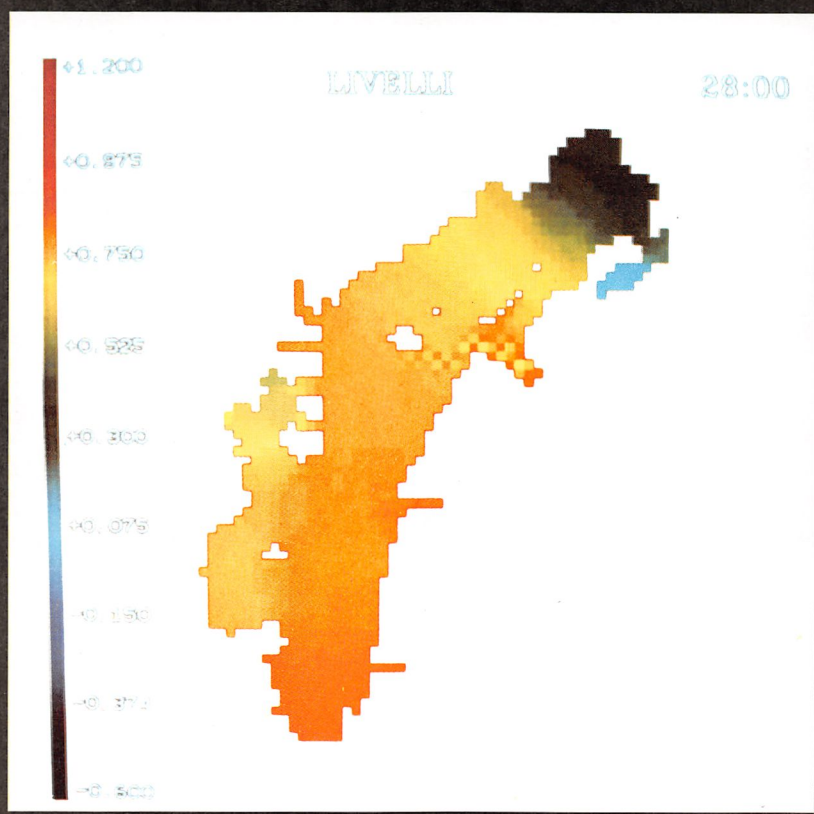
3



3. Diagram showing the way a computer would execute the instruction LOAD A.

The kind of information which can be generated by the use of computers is enormously varied. The images on p. 133 obtained by using the HACIENDA system, were produced to illustrate the evolution of tides in the Venice Lagoon. Different colours show different sea levels (expressed with respect to mean sea level) as indicated on the chromatic scale to the left of the photo. (photo: courtesy IBM).

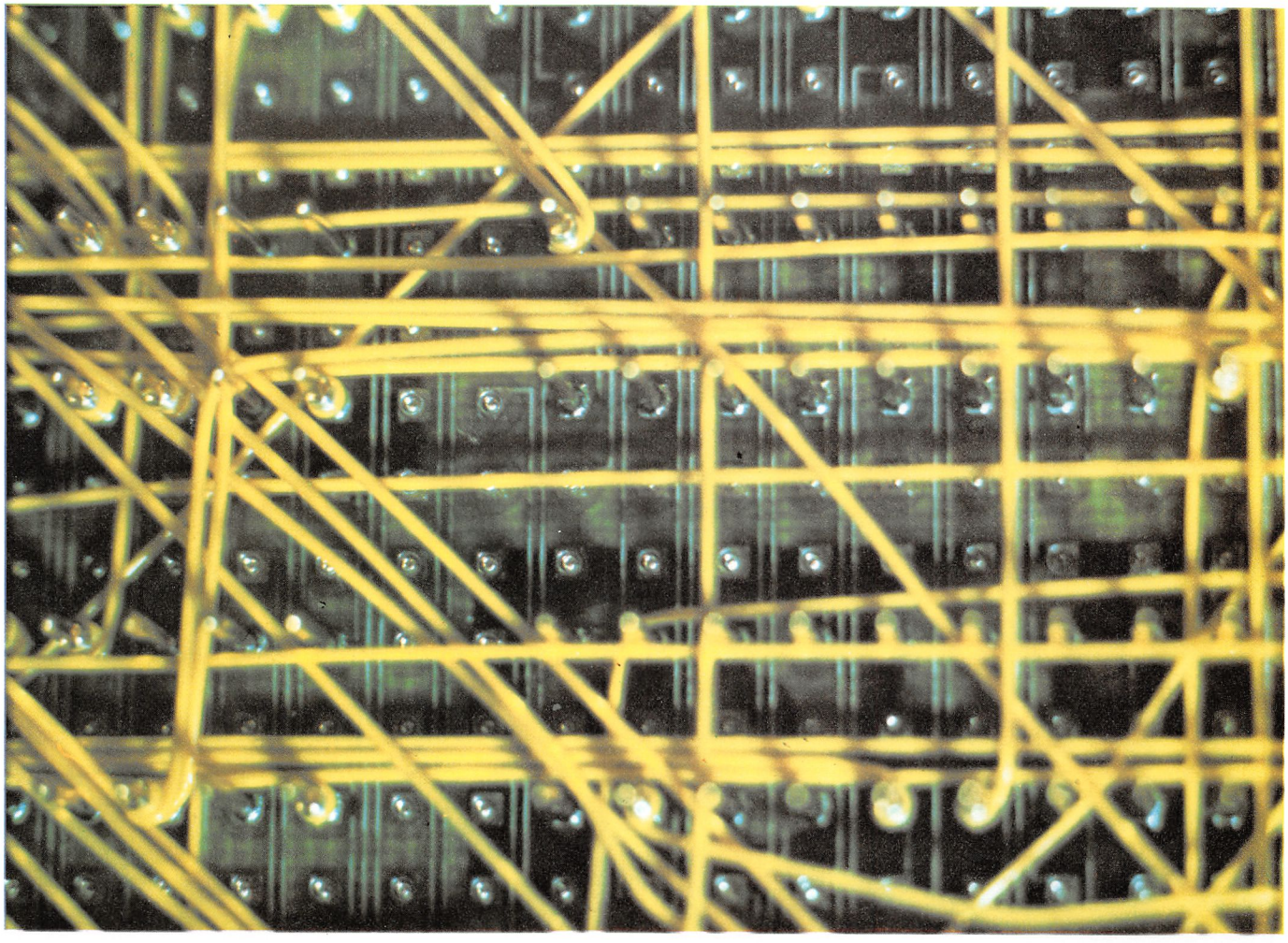

+1.200
+0.875
+0.750
+0.525
+0.300
+0.075
-0.130
-0.374
-0.600
LIVELLI
27:00


+1.200
+0.875
+0.750
+0.525
+0.300
+0.075
-0.130
-0.374
-0.600
LIVELLI
28:00

the address of the operation code of the instruction, is put on the address bus and the program counter is increased by 1, so that the address of the next instruction can be found;
2) the signals of the control bus are activated. They indicate that there is a reading operation from memory of an operation code (**fetch**);
3) the memory receives the address present on the address bus and the signals of the control bus as input. It answers by putting the contents of the word (a byte, in this case) which it found at that address, on the data bus;
4) the CPU receives the byte read from memory on the data bus. As it is an operation code it is loaded into a special register of the CPU itself, where it is decoded. This is called the **instruction register** (IR);
5) the code 3E is interpreted as an operation of reading from memory;
6) to read the following byte which contains the operand 0F, the preceding operations are repeated with these differences: the control signals for a *memory read* are activated, but not those which signal a fetch operation, because in this case, the byte read contains data and not an operation code. The byte read is loaded into the **accumulator** (because it is not an instruction), as required by the instruction under way.

To recap on the overall process: the computer initially requires a program within its memory. The complete program must be in sequence and the address of the first instruction of the program is put into the program counter.

In early computers this **loading** of the program and first instruction address were lengthy operations because each instruction of the program was loaded with the



In the photo below you can see part of the tangled mass of wires used to connect a large number of logic circuits in an old IBM computer. Comparing this with any modern computer circuit board illustrates how relatively streamlined hardware has now become, with high levels of circuit integration. (photo IBM).

use of a number of mechanical switches.

Each switch represented one bit of a binary instruction and its proposed memory address and, when all switches had been set 'on' or 'off' (i.e. to logic 1, or logic 0) according to the corresponding bit of the instruction and address, a separate switch was then operated to load the instruction into the specified memory address.

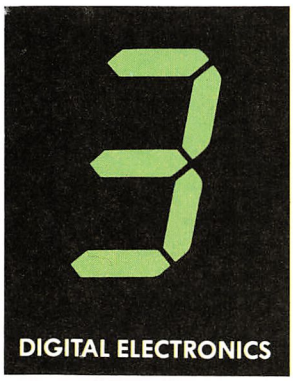
After the whole program had been loaded this way, instruction by instruction into memory, the switches were then used to load the address of the first instruction into the program counter. Only at this point could the program be run and the computer made operational.

In modern computers, like all personal or home computers, all this happens automatically after turning on. In such machines, turning on causes a 'start-up' or **initialisation** program to occur, which rapidly (within a second or so) makes the computer operational. Chapters covering basic software and operating systems will explain this in detail.

The majority of this type of computer provides the user with a keyboard and a TV-type screen for communication, but direct or manual access to particular addresses in memory or the program counter is not usually possible. Some computers, however, allow a user to access individual addresses with a special type of program instruction.

Glossary

BCD	binary coded decimal representation – a system which uses binary digits to represent decimal numbers. Each decimal digit is represented by four binary digits
binary system	number system arranged according to the base 2. It uses the digits 0 and 1
double precision	system where floating point numbers are represented in twice the normal number of words. This allows a greater accuracy in calculations
external alphabet	all the characters that a computer can read in or print out, usually composed of the 26 letters of the alphabet, the ten decimal digits, graphics characters, operational symbols and punctuation marks
floating point representation	system where numbers are represented as a fraction (mantissa) multiplied by the base number to a power (exponent). For instance the decimal number -385.89 is represented as -0.38589×10^3 and the binary number 110111.01101 as 0.11011101101×2^6
hexadecimal system	number system arranged according to the base 16. It uses the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F
machine language	the codes which represent the computer's operating instructions. These are directly understood by the machine
octal system	number system arranged according to the base 8. It uses the digits 0, 1, 2, 3, 4, 5, 6, 7
single precision	the normal method used to represent floating point numbers in a computer



Integrated logic circuits

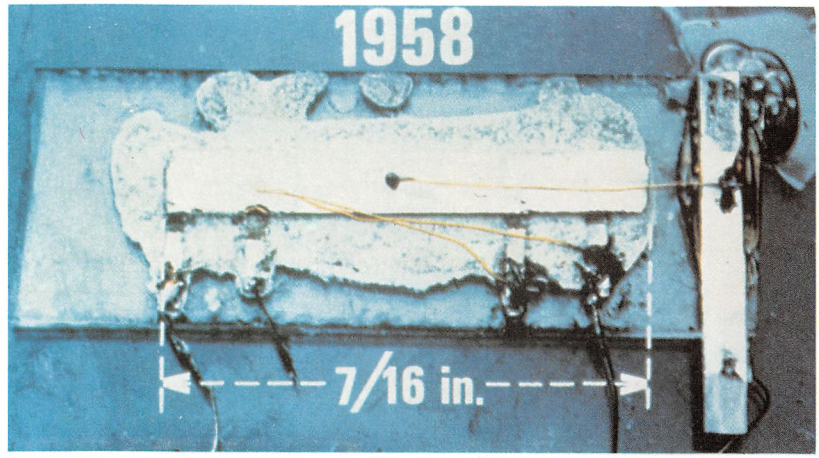
Development of integration

In the early days of computer design all the circuits were constructed using **discrete** components, first thermionic valves and later transistors. Even transistor circuits were slow by present-day standards and they also consumed a large amount of power. Nowadays **integrated** circuits are used. These can be produced much more cheaply, consume less power and operate much faster.

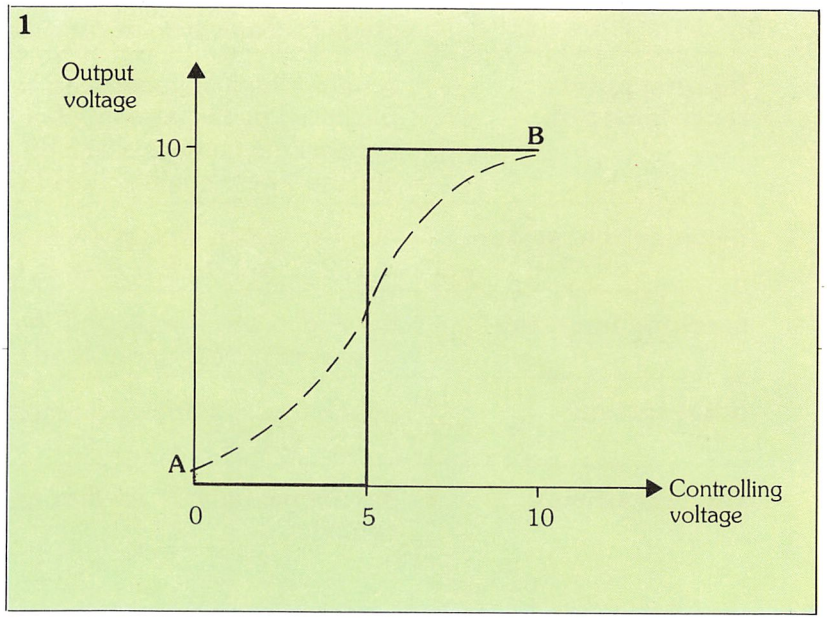
The basic idea of using logic gates made of electronic circuit elements has been introduced in earlier chapters, with examples using NMOS and bipolar transistors. We now need to look at the various families of **integrated** logic circuits and their suitability for various applications. Each family group consists of NOT, OR, NOR, AND, NAND gates, all of which are constructed using one type of technology. Probably the two most common families are **TTL** (transistor-transistor logic) and **CMOS** (complementary metal oxide semiconductor).

Other family groups include the earlier and now obsolete **DCTL** (direct coupled transistor logic), **RTL** (resistor transistor logic) **DTL** (diode transistor logic), and also the more modern families such as **NMOS** (n-channel MOS), **ECL** (emitter coupled logic) and **I^2L** (integrated injection logic).

As mentioned above, the basic gates NOT, OR, NOR, AND, NAND may be constructed using any of these devices. In more advanced circuits, large numbers of these devices have to be interconnected to make as compact a circuit as possible. Designers will choose the circuits which allow them to meet the specification's requirements with the smallest overall cost. This inevitably involves compromise: higher performance may be achieved, but for an increased cost. It's a question of choosing the right circuits for the job. A pocket calculator which has to be competitively priced may use a design which operates relatively slowly, as a wait of one or two seconds for an answer is not usually very serious. However a large computer requires circuits which will give answers in less than 1 microsecond ($\mu s = 1 \times 10^{-6} s$) since many calculations need to be performed very rapidly one after the other.



The first integrated circuit made by Texas Instruments in 1958 (US patent 3, 138, 743).



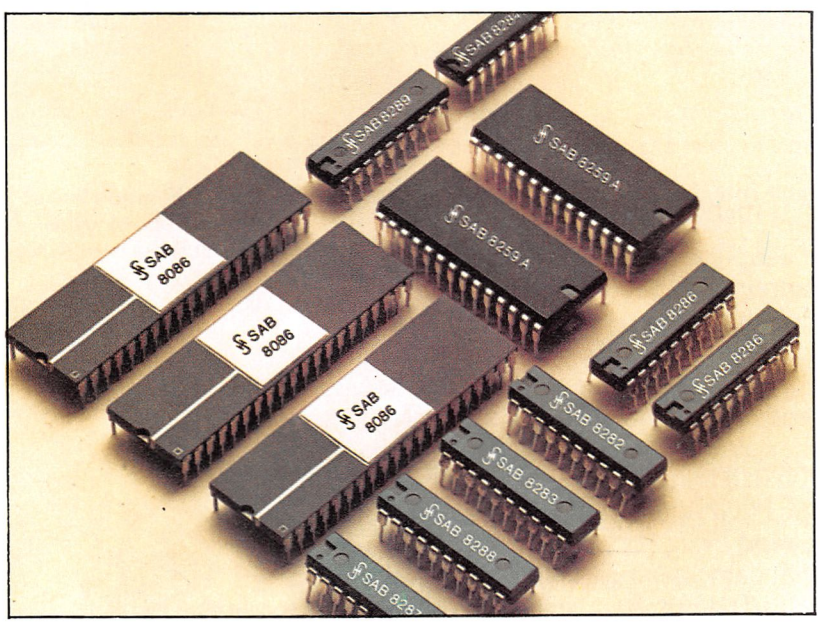
1. Graph showing the transition between off and on in a logic gate. The dotted line shows the transfer characteristic of the device.

Characteristics of logic circuits

Up to now we have used a simple switch, operated by a specified voltage, to represent a gate. This operation may be described by the graph of *figure 1*. Point A shows that when the controlling voltage is 0 V, the output is disconnected from the supply and is thus 0 V. When the controlling voltage is 10 volts, the switch is closed and the output is connected to the supply of 10 volts, point B. We have said nothing about what happens if the controlling voltage is say, 5 volts, or anywhere between the two extreme values. Ideally, for the circuit to clearly differentiate and switch between these two values, the control voltage should have a direct effect on the output, as shown by the solid line in *figure 1*. Here the switch closes precisely when the control voltage reaches a fixed value, in our case, 5 V.

In practice no electronic circuit will behave just like this. The transition between the non-conducting and conducting states will be more gradual, looking something like the broken line in *figure 1*. It is unusual for the non-conducting condition to correspond exactly to 0 V, so in general the **LOW** state is defined as logic 0 (using positive logic). Obviously the closer this is to 0 V the better the device. Logic 1 is defined as the **HIGH** state – again this will probably not be identical to the supply voltage.

Different types of integrated circuits (supplied by Seimens).



As you will have noticed a single gate can only perform very simple logic functions. To perform more complex operations gates have to be connected together. This is done by connecting the output of one gate to one of the inputs of a subsequent gate. This operation is known as **cascading.** The signals applied to the inputs of a complete network will set the outputs of the first line of gates to the appropriate values. These in turn will feed the second line of gates and so on, until the last gate is reached, where the output of the complete system is obtained.

It is rather like knocking down the first of a line of dominoes: each one knocks down the next one (or two besides each other) and this operation ripples through the system. The term **ripple** is used to describe the way information is transmitted through some of the simpler types of computing system. Following the analogy further, each domino requires a small but definite time before hitting the next; in a similar manner each gate takes a short time after receiving its control signal before it actually operates its switch. This is known as the **operating speed.**

Let's now define some of the important properties of gates, which determine the way they are used by a circuit designer.

Threshold voltage

This is a rather arbitrary voltage level at which the gate switches from low to high. In *figure 1* it is assumed to be the average of the two extreme values (i.e. 5 V). Thus for any controlling voltage below 5 V the output is in the LOW state and for voltages above 5 V it is in the HIGH state. This is fine if the **transfer characteristic** is represented by the solid line in *figure 1* but in reality the transfer characteristic is like that shown by the dotted line. It is obvious that we cannot define 5 volts as being the threshold voltage as it does not correspond to a definite state of switching.

So it is important to define some logic **bounds** (ranges of permitted voltages) on LOW and HIGH states. *Figure 2* illustrates that any voltage between 0 and 1 V can be defined as being in the LOW state and

anything between 8 and 10 V as being in the HIGH state. (The values of these bounds are left to the designer of the device.)

For a LOW control voltage between 0 and 1 V, the output voltage lies between 0.3 and 0.5 V and is well within the bounds for logic 0 (LOW). Similarly for a HIGH control voltage between 8 and 10 V, the output will lie between 9 and 9.3 V and so is within the bounds of logic 1 (HIGH).

Noise margin

Let's consider a LOW control voltage at its highest permitted value of 1 V, and suppose we have some undesired voltage added to it. These additional undesired voltages are termed **noise,** and may be caused by interference from power lines, X rays, lightning, car ignition, and so on. Now if this noise voltage is less than 1.3 V, the total control voltage is 2.3 V. We can see from *figure 2* that the output voltage will be less than 1 V and hence still in the LOW state. If the noise voltage is greater than 1.3 V then the output will be above 1 V. The circuit will then not give a correct output and is therefore no longer working satisfactorily. In this circuit the **noise margin** in the LOW state is 1.3 V.

Similarly a HIGH control voltage of 8 V together with a noise voltage greater than – 2.3 V will give us a total voltage of less than 5.7 V. The output will be less than 8 V and again the circuit will not work correctly. The **noise margin** in the HIGH state is thus 2.3 V.

Operating speed

Changes of state in a gate are a result of changes in the quantity of charge stored in the devices the gate is made from. Like the domino analogy, each gate requires a certain time to operate and this is termed the **propagation delay time** of the gate. When a number of gates are connected in cascade the total delay is the sum of the delays in all the cascaded elements. Propagation delay times lie in the range 2 to 50 nanoseconds (ns = 1×10^{-9}s). The time to change from LOW to HIGH is usually different to the time from HIGH to LOW.

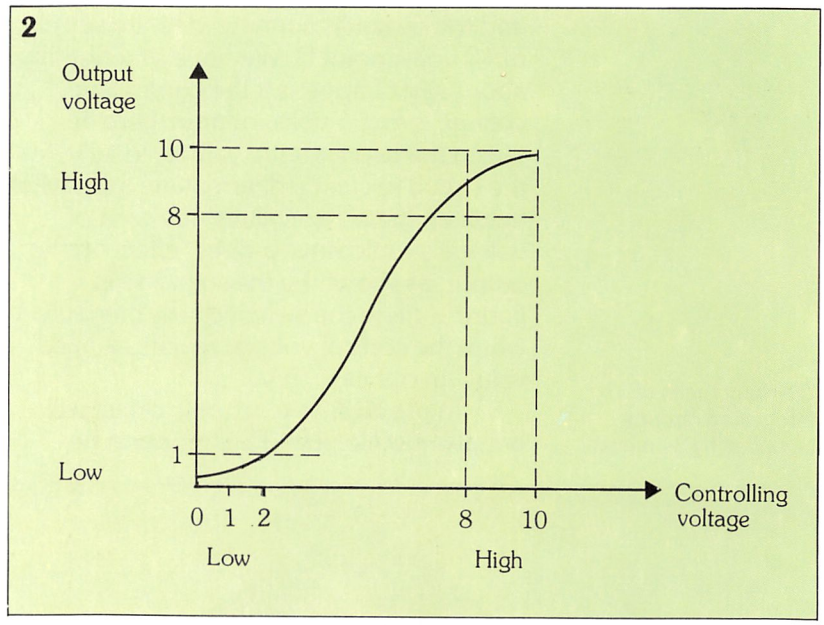
Fan-out and fan-in

Most of the gates which we have met have more than one input terminal. For example, a gate having three input terminals may be connected to three input controlling signals. It is said to have a fan-in of three. The **fan-in** is thus the number of input terminals on a gate.

We have said that the logic levels at the output of a circuit are restricted to two bands: HIGH and LOW. These are usually defined with the output of the circuit left unconnected to any load. When circuits are cascaded, the output of any chosen gate may be connected to the input terminals of one or more other gates. Since the inputs of every gate require a certain current, this needs to be supplied by the output of the preceding gate. The load current of the output will have to be able to supply all the inputs connected to it. As the load current drawn from a gate increases, the value of the output voltage will start to rise (if it is in the LOW state) or fall (if it is in the HIGH state) until it comes out of the **bound** (range of permitted voltages) set for that particular logic state.

The number of input control gates which may be connected to the output of a preceding gate is the **fan-out** of the gate.

2



2. Dotted graph line of figure 1, showing the high and low bounds set for the controlling voltage.

Power dissipation

In an ideal gate, when no current flows through it the voltage across the gate is equal to the supply voltage. Since the power dissipated is equal to the voltage multiplied by the current, no power is lost.

Similarly when the gate is closed and current flows the voltage across the gate is 0 V and still no power is lost. However, practical gates cannot achieve these ideals. There will always be some current flow through, or some voltage across, the gate.

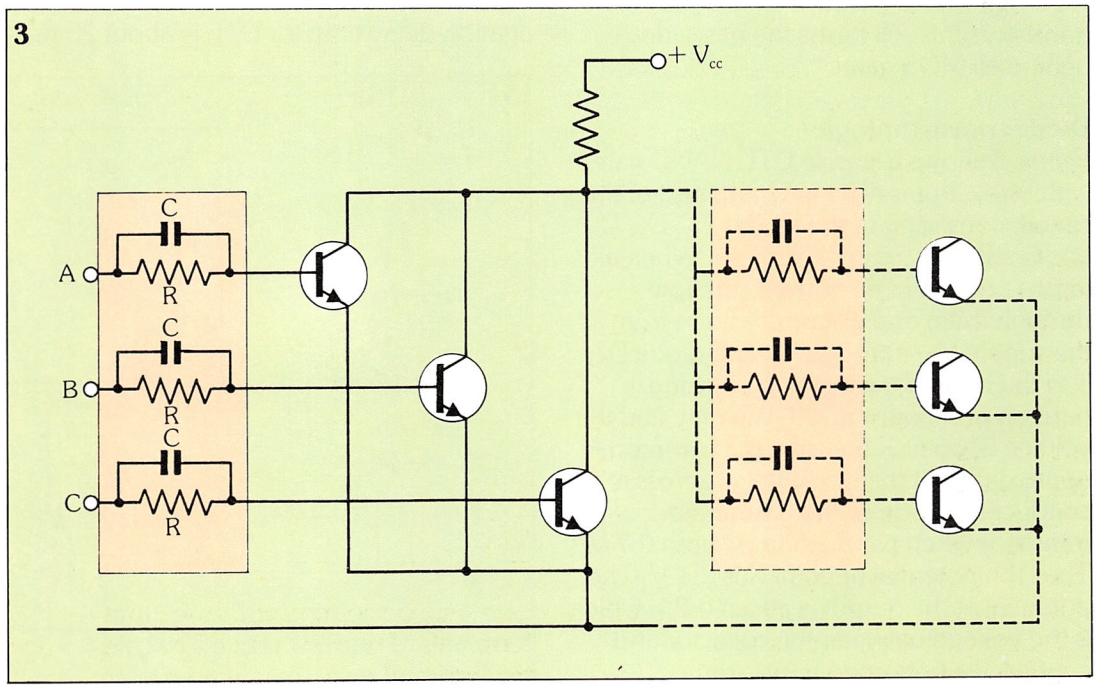
The power dissipated in a gate is the average of the power dissipated in HIGH and LOW states. This gives an indication of the amount of heat which must be transferred to the surroundings to avoid excessive generation of heat and subsequent damage to the gate.

Logic families

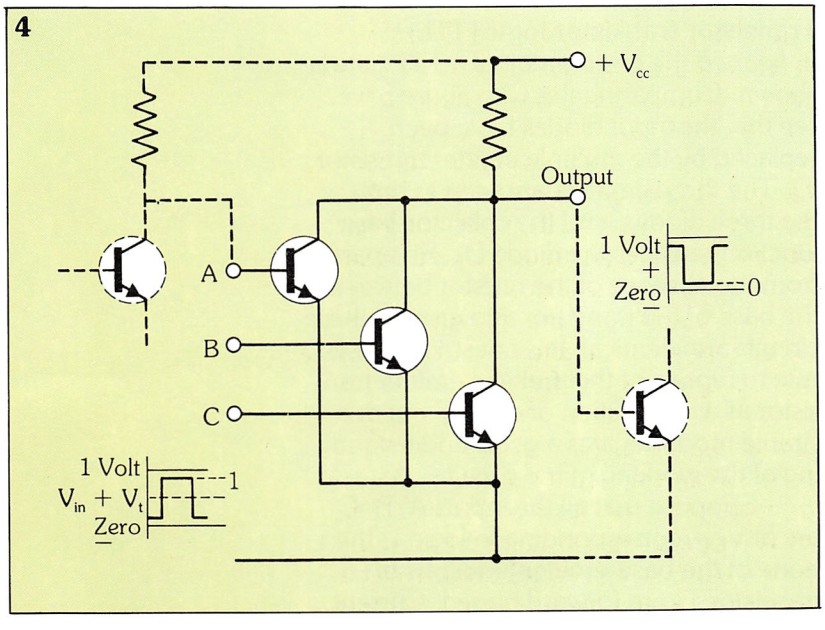
Looking now at different families of logic circuits, we will briefly cover older families and concentrate in more detail on those used in modern integrated circuits.

Resistor transistor logic (RTL)
This was the first of the logic families and was mainly used with discrete transistor circuits. The basic arrangement is shown in *figure 3*. The transistors are switched on by raising the inputs A, B or C to the positive supply voltage V_{CC}, which is approximately 5 V. The base current flowing through the resistors R causes the transistors to bottom, giving an output voltage of about 0.2 V. The capacitors are included so that the transition from OFF to ON may be made faster. The transistors are switched OFF by returning the input terminals to 0 V thereby stopping the flow of collector current. The fan-out of RTL is low and the noise margins are small.

3. The basic circuit arrangement of an RTL NOR gate.



4. The basic circuit arrangement of a DCTL NAND gate.



Direct coupled transistor logic (DCTL)
A modification of RTL in which the input resistors are omitted is shown in *figure 4*, and is known as Direct Coupled Transistor Logic. This arrangement has the advantage that switching is faster, however, the difference between the HIGH and LOW

states is severely limited. When the transistors are conducting in the bottomed state, the voltage on their collectors is about 0.2 V. When they are OFF the voltage cannot rise above about 0.8 V, since the maximum base voltage of the following transistor is limited to this value. The advantages of the circuit are simplicity and fast switching speed, but the noise immunity is very poor. Another disadvantage is that if the transistors are not absolutely identical one will 'hog' more of the current than the others, causing the circuit to malfunction. Although no longer used with bipolar transistors this still forms the basic idea behind all MOS gates.

Diode transistor logic

Figure 5 shows a simple DTL NAND gate with three inputs. The left hand part of this circuit, consisting of the diodes D_1, D_2, D_3 and resistor R_D, is an AND gate. When all inputs are HIGH no current can flow through them and all current flows from the supply V_{CC} through R_D and diode D_S, into the base of transistor T_1, turning it hard on and bottoming it. We may find the voltage at point X by using the common approximation that the voltage across a conducting diode or the base-emitter of a transistor when conducting is about 0.7 V. Thus the potential at point X is 2.1 V. The potential at the output is about 0.2 V which is the voltage between the collector and emitter of a bottomed transistor.

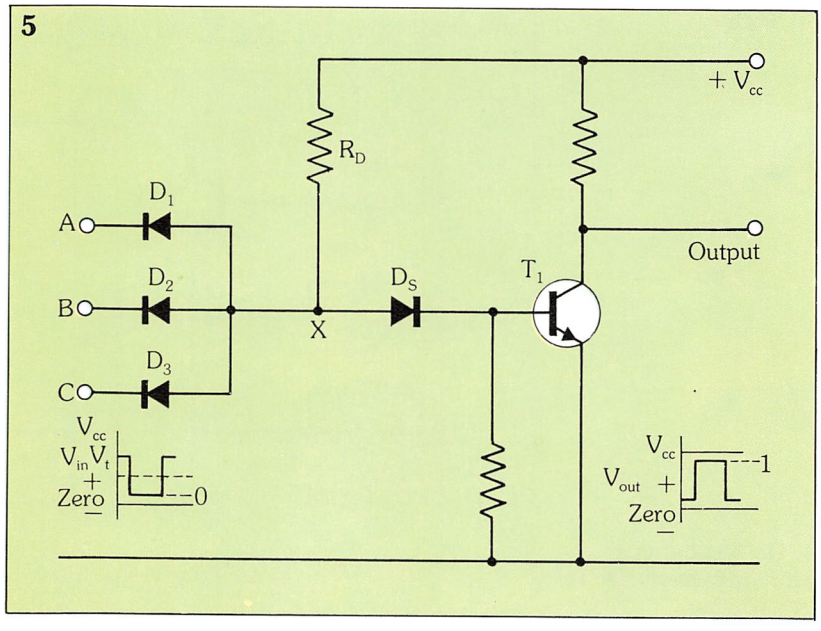
If any of the input terminals A, B or C is connected to 0 V this allows current to flow through R_D and the relevant diode. The potential of X now falls to 0.7 V. This voltage is too low to allow sufficient current to flow through D_S into the base of T_1 to cause it to conduct, and the output of the circuit will rise to approximately V_{CC}, corresponding to logic 1.

As the voltage across a conducting diode is 0.7 V, any noise voltage equal to or greater than this will be regarded by the circuit as an input signal. This noise margin is relatively small, and may be improved by replacing the single diode D_S by two diodes in series. The voltage at X when T_1 is conducting is now 2.8 V, with a noise margin of 1.4 V. However we cannot keep adding diodes indefinitely, as the addition of extra diodes makes the circuits work more slowly, increasing the propagation delay time.

Diodes are used on the inputs of these devices (instead of resistors as in an RTL circuit) as semiconductors are more easily constructed and take up less space on integrated circuits than resistors and capacitors.

The speed of operation of DTL is faster than that of RTL since the input signal has to flow through the very low resistance of the diodes rather than through input resistors, and no parallel capacitors are required. The typical propagation delay time for DTL is about 25 ns.

5. Basic circuit arrangement of a DTL NAND gate.



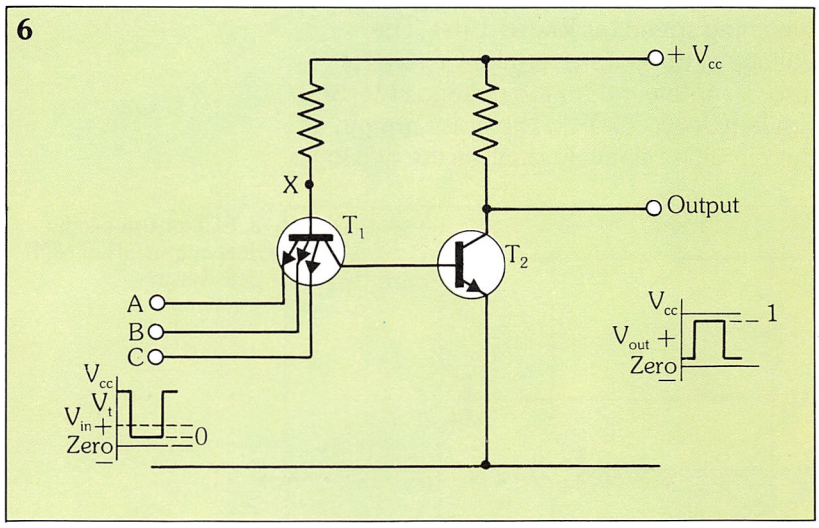
Transistor transistor logic (TTL)

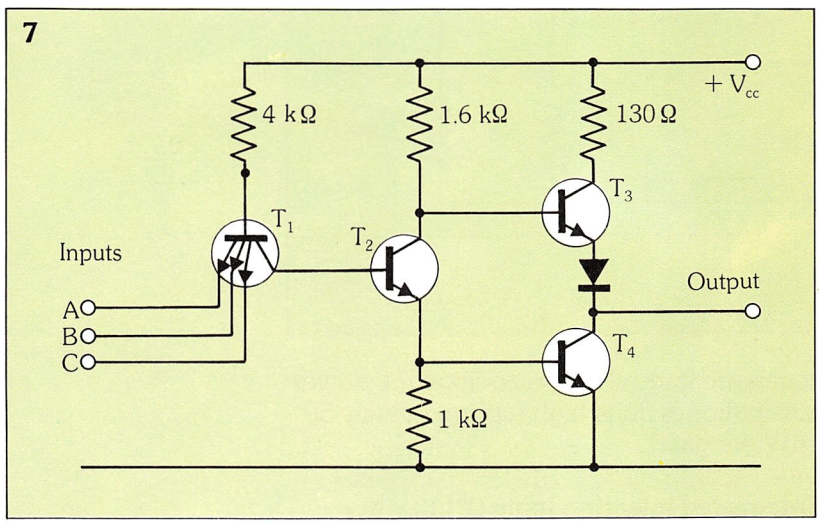
In *figure 6* the basic circuit of a TTL gate is shown. Comparing this with *figure 5* we see that the input diodes have been replaced by the multiple emitter transistor T_1. The three emitters are seen to replace the three diodes, and the collector-base junction replaces the diode D_S. So apart from the absence of the resistor between the base of the transistor and ground these circuits are identical. It is not strictly accurate to represent the multiple emitter transistor as a connection of diodes, but as a simple model it gives a good understanding of the working of the circuit.

Suppose that all the inputs A, B, C are at V_{CC}, corresponding to logic 1, then none of the base-emitter junctions of transistor T_1 are forward biased. Current

flows through it and into the base of T_2, switching it hard ON and bottoming it, forcing the output to about 0.2 V. In order to keep T_2 bottomed, it is necessary for point X to be at 1.4 V as we discussed for DTL. If one of the emitters of T_1 is now connected to 0 V, current will flow through this emitter (its value being limited by the resistor in series with the base), and point X will be pulled down to about 0.7 V.

6. Basic TTL NAND gate circuit.





7. Practical circuit arrangement of a TTL NAND gate, using a totem-pole output stage.

This will not be large enough to keep transistor T_2 conducting, so the output voltage will rise to V_{CC}, corresponding to logic 1.

One great advantage of TTL is that it is considerably simpler to manufacture one multiple emitter transistor than several diodes.

The circuit of *figure 6* is somewhat limited as it has a very small fan-out. If we try cascading this circuit with a number of other gates, the current that each of these forces into the output terminal when it is in the LOW state will cause transistor T_2 to cease to be bottomed. The circuit will then no longer operate satisfactorily.

This may be remedied as in *figure 7* by adding a power output stage often termed a **totem-pole.** When transistor T_2 is switched OFF, its collector will rise towards V_{CC} and its emitter will fall to 0 V. This causes the base-emitter juncion of T_3 to be forward biased. T_3 will conduct and the voltage between its collector and emitter will be held at about 0.2 V, causing the output voltage to be high. At the same time the base of transistor T_4 will be at 0 V, so the transistor will not conduct and its collector-emitter voltage will be large.

When T_2 is conducting, its emitter will be at a positive voltage and this will cause transistor T_4 to conduct and become bottomed, holding the output at 0.2 V. Also the base-emitter junction of T_3 will be reverse biased and so it will be non-conducting. There are thus two low resistance paths for the output current, either through T_4 when the output is LOW or through T_3 when the output is HIGH. This increases the fan-out and reduces the propagation delay time. We may determine the voltage at X when the output is LOW as the sum of the voltage drops across the forward biased emitter-base junctions of T_2 and T_4 and the forward biased collector-base junction of T_1; it is therefore 2.1 V. If A, B or C is set to 0 V the voltage at X is 0.7 V. We thus have a noise margin of about 1.4 V for this circuit.

The TTL family of circuits is designed to operate from a supply voltage of about 5 V. This allows it to be used directly with other families (particularly DTL), with no need for any special circuits to interface one type with the other.

Variations on the basic TTL family give rise to sub families which either are designed to have **low power dissipation** or alternatively to have a **faster operating time.** The main limitation on the speed of operation of a gate results from the need to inject or remove charge from the transistors in the gate. Whenever a transistor is bottomed, large amounts of charge are forced into its base. The harder it is

bottomed the more charge is stored. A sub group of TTL gates, **Schottky TTL** overcomes this problem by making sure that the transistors are never quite bottomed.

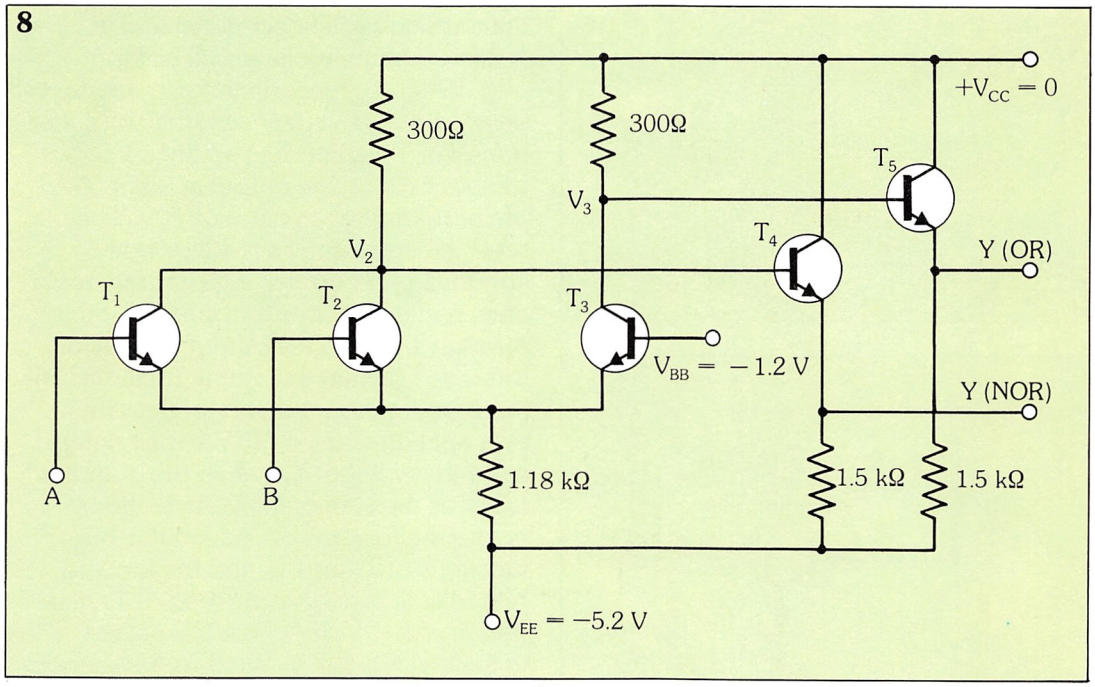
TTL is the basic form of logic used for small scale integration, where only a few gates are used on one chip. They are cheap, they have relatively low power dissipation, a high fan-out and are reasonably fast.

Emitter coupled logic (ECL)

Another form of logic which does not allow the transistors to bottom is shown in *figure 8*, which is a positive logic OR or NOR gate. This type of design allows one circuit to perform both functions at different output terminals. If both inputs are at logic 0 (in this circuit −1.7 V) both transistors T_1 and T_2 are non-conducting. As V_{BB} is fixed at −1.2 V, transistor T_3 conducts and the voltage at point X is approximately −1.9 V (0.7 V below the base voltage). Since T_1 and T_2 are non-conducting, V_2 is approximately at 0 volts and the current through T_3 sets V_3 at about −1.0 V (the LOW voltage level for this circuit).

If A is now raised to the HIGH level (−0.7 V) transistor T_1 will conduct. The voltage at X will rise to −1.4 V and the voltage at V_2 will be about −1.0 V. The base emitter junction of transistor T_3 will now be reverse biased and it will no longer conduct, so the voltage at V_2 will rise to 0 V.

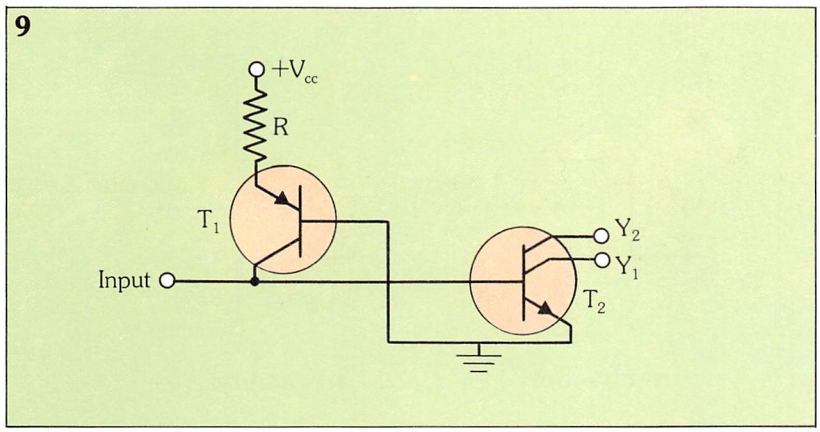
To sum up: if both A and B are at logic 0 (−1.7 V), V_2 is at 0 volts (HIGH) and V_3 is at −1.2 V (LOW). If either A or B, or both, are at logic 1 (−0.7 V) then V_2 will be at logic 0 and V_3 will give a NOR output. The addition of the two transistors T_4 and T_5 at the output allow considerably larger currents to flow, increasing the fan-out (up to 30) and improving the switching speed (as low as 1 ns). The voltage at the output terminal Y (OR) is also 0.7 V lower than the voltage at V_3 and similarly for Y (NOR). The noise margin of this circuit is relatively small as the two logic levels are fairly close. In addition the power dissipation is quite high (approximately 50 mW per gate).



8. ECL positive logic dual output OR or NOR gate design.

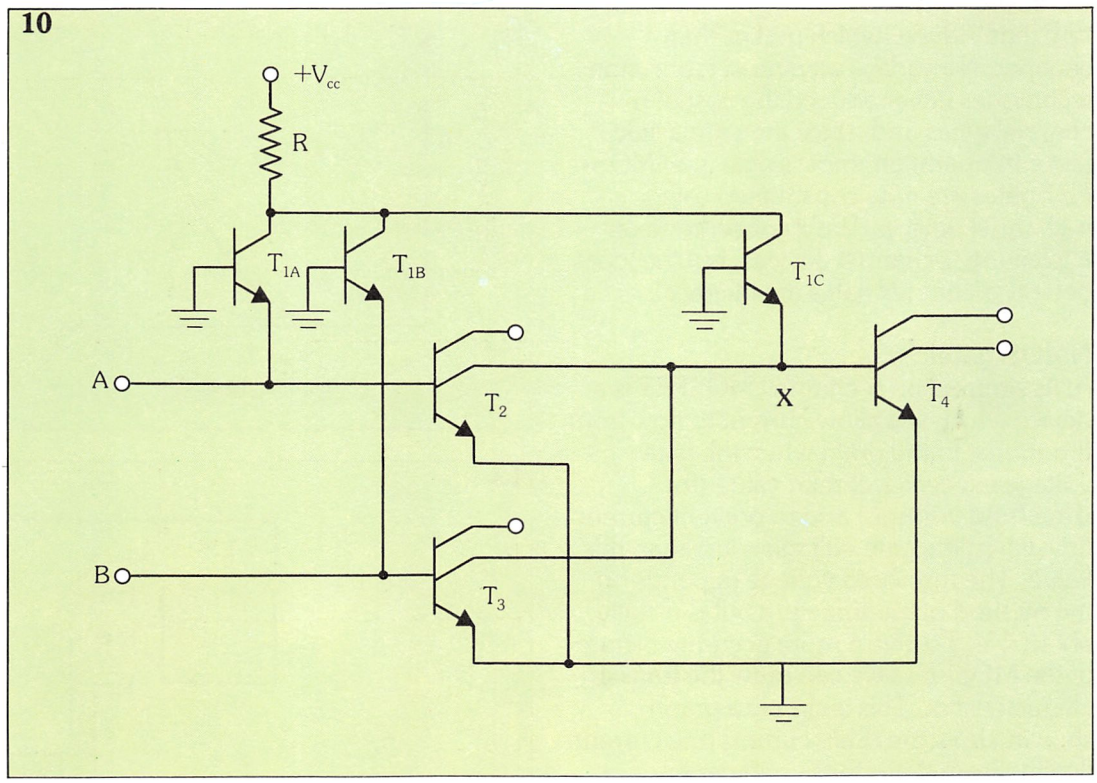
Integrated injection logic (I^2L)

This is one of the newest types of logic family. Its basic form is shown in *figure 9*. It is very similar in operation to DCTL but a number of the transistors have been amalgamated to form multiple collector devices. If the input terminal in *figure 9* is a short circuit to ground (0 V), logic 0, all the current through transistor T_1 will flow through the short circuit and transistor T_2 will be switched OFF. Thus no current could flow in any of the collectors of T_2



9. Basic circuit design of a I^2L gate using a multiple collector transistor.

10. I^2L design for a positive OR gate.



To construct a positive OR gate we use the circuit shown in *figure 10*. Two I^2L circuits are connected in parallel to the base of T_4. If A or B is at logic 1 (0.7 V) then transistor T_2 or transistor T_3 conducts, and point X will appear shorted to ground. If both A and B are at logic 0 (0 V) transistors T_2 and T_3 cannot conduct and point X will appear as an open circuit. This gives the function of an OR gate. Many such gates can be cascaded one after another.

One of the great advantages of the I^2L family is that no resistors are used (the single resistor R is external to the chip and is common to all the gates in any complex circuit); I^2L ICs are therefore easy to make and cheap. At low speeds very little power is consumed (5 nW) and even at high speeds power consumption is only about 5 mW. Principal applications are in digital watches and small microprocessors.

which are then open circuits.

On the other hand if the input is left on open circuit so that no current can flow, all the current through transistor T_1 will flow into the base of T_2 and switch it full on (bottomed). The collectors of T_2 will then appear to be short-circuited to ground and they will be at 0 V. In this condition the input terminal will be held at 0.7 V (the base emitter voltage of T_2 cannot be greater than 0.7 V). The output terminals of this circuit can represent either a short-circuit (0 V) or an open circuit (0.7 V) depending on whether the input terminals are open circuit or short circuit to ground.

MOS circuits

In chapter 2 we looked at some simple gates using MOSFETS. There are a number of logic families based on these, principally PMOS, (using p-channel MOSFETS), NMOS (using the n-channel de-

vices which we have briefly covered) and CMOS which uses a combination of both types. The symbols for the two types of MOSFET are given in *figure 11* which shows both p and n-channel enhancement types. Note that the only difference between these two symbols is the direction of the arrow on the substrate. The substrate is the basic chip of silicon on which many MOSFETS are constructed, and the source terminal is frequently connected to this in practical circuits.

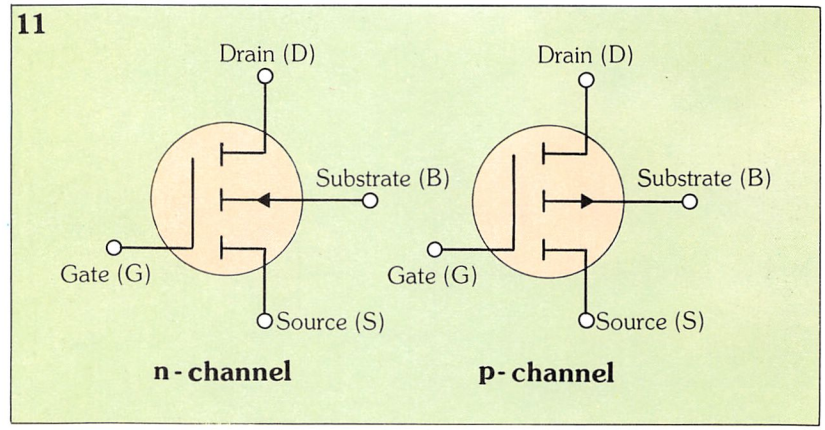
The earliest MOSFET logic gates were constructed using p-channel devices. These are somewhat simpler to manufacture than n-type, which makes them cheaper. Nowadays advanced fabrication techniques have reduced the cost of n-channel types and, since they are a little faster in operation, most single type MOSFET gates are now constructed using n-channel devices. We'll concentrate on explaining n-channel devices, but the ideas behind p-channel gates are identical.

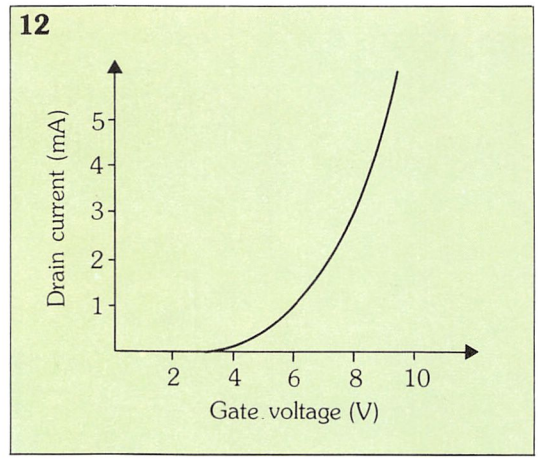
NMOS gates

At its simplest an n-channel MOSFET is a device which will allow current to flow from the source to the drain when the gate voltage exceeds a certain value (the **threshold voltage**) and to prevent current flow when the gate voltage is less than this value. The threshold voltage depends on the method of manufacture but is usually about 4 V. To give a more accurate picture of the MOSFETs we can draw the transfer characteristics. This is simply a graph showing how the drain current (the current flowing from drain to source) depends on the gate voltage, and is drawn in *figure 12*.

One of the most usual circuits using MOSFET is shown in *figure 13*. This is a simple inverter (NOT) gate. Transistor T_2 has its gate connected directly to its drain. It works like a simple resistor, although the current is no longer directly proportional to the voltage. The reason why MOSFETs are used instead of resistors is that they are smaller and easier to make on ICs.

We can now see that transistor T_1 will be an open circuit if the input voltage is less than V_{th} (the threshold voltage) and thus the output voltage will be equal to 10 V. If the output voltage is greater than V_{th} current will flow and the output voltage will fall to about 1 V. The more exact relationship is shown by the graph of *figure 14*. We can see that the noise margin of MOSFET circuits is large (about 3 V) – which is quite an advantage over other logic families.

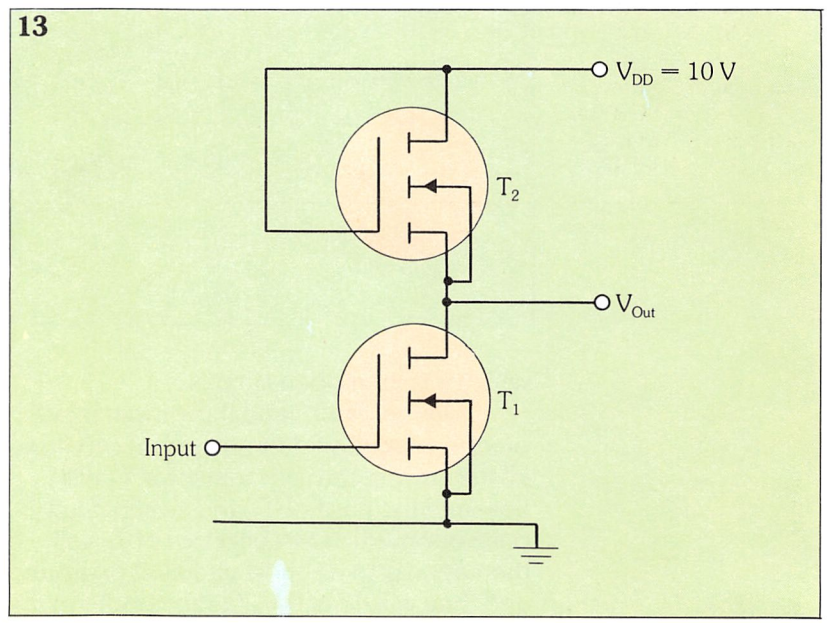


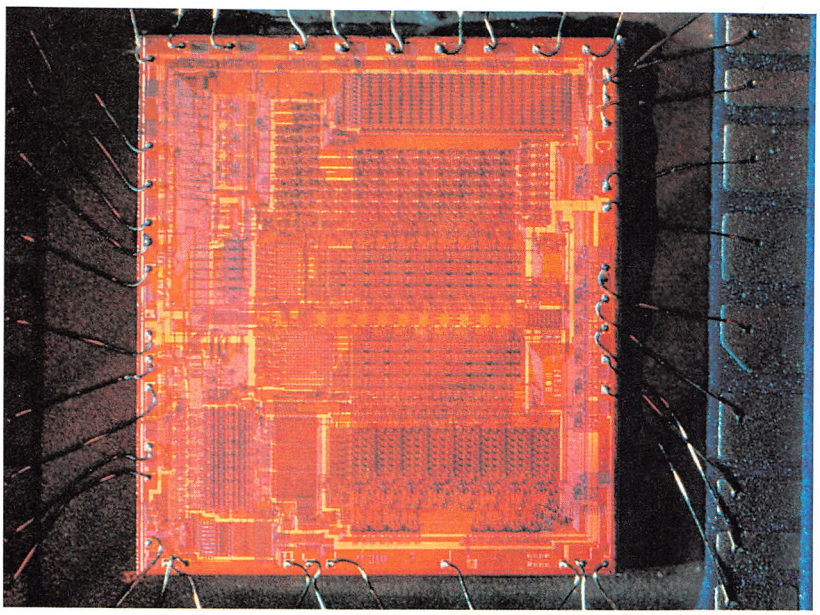


11. An n- and p-channel MOSFET.

12. Graph showing the relation between gate voltage and drain current in a MOSFET.

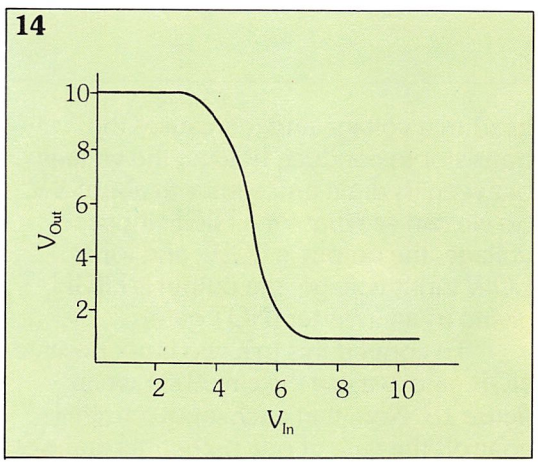
13. A simple inverter (NOT) gate, made from n-channel MOSFETs.





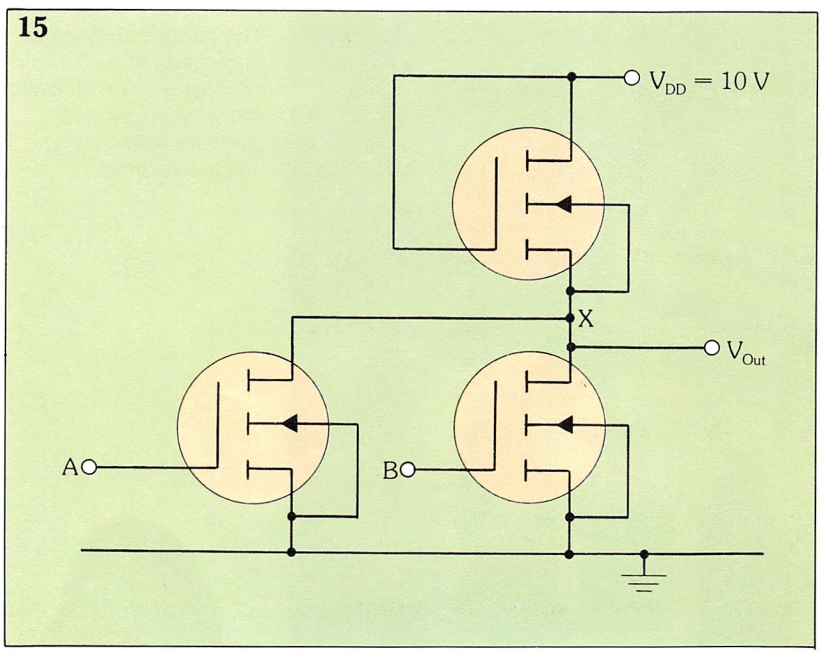
Science Photo Library

MOS technology microprocessor IC, highly magnified.



14. Relationship of input and output voltages for the circuit in figure 13.

15. Circuit of figure 13 converted to a NOR gate by the addition of third n-channel MOSFET.



Another advantage is that the gate current of a MOSFET is extremely small (typically as small as one picoamp or even less) whether it is conducting or not. So many gates may be cascaded without excessive loading. In fact the number of gates which can be connected together to the output is almost limitless and we have a large fan-out capacity. However this is limited by the reduction in speed as the loading is increased.

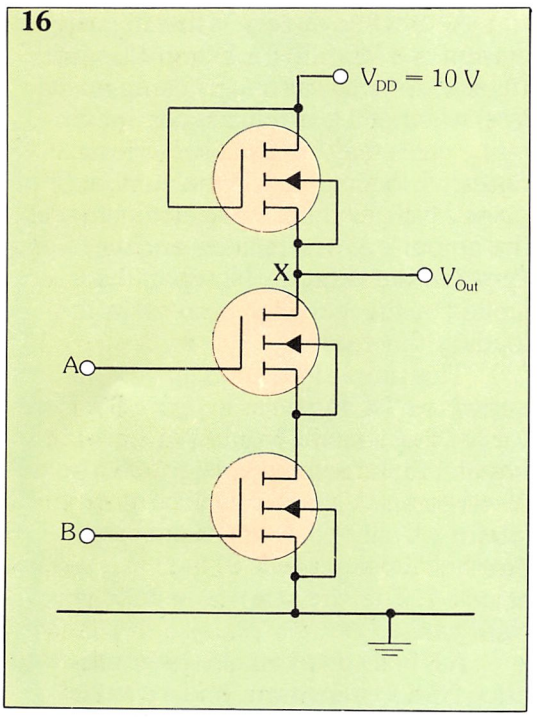
This simple inverter may be converted to a NOR gate as in *figure 15*. Here we see that if either A or B is at logic 1 (greater than theshold voltage) then point X will be at 0 V and V_{out} will be at logic 0 (about 1 V). If both A and B are below threshold voltage (logic 0) then V_{out} will be at logic 1. This circuit is now a positive NOR gate.

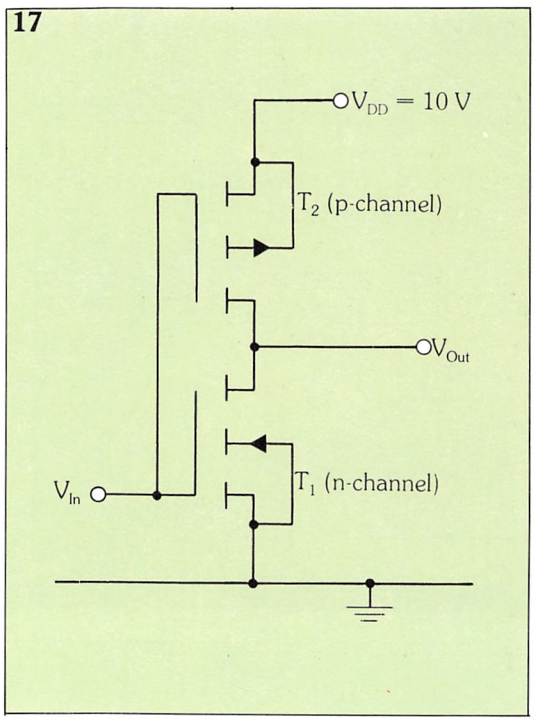
A NAND gate can also be constructed using NMOS transistors, and a circuit is shown in *figure 16*. If input A is logic 0, the transistor does not conduct and V_{out} must be at logic 1. Similarly, V_{out} will be at logic 1 if input B is at logic 0. However, if both inputs are at logic 1 then both transistors conduct and V_{out} must be at logic 0. This, of course, is the positive logic NAND function.

NMOS logic, although slower than TTL, has a packing density (the number of individual logic gates per area) that is very much higher than TTL and is cheaper to make. For these and other reasons it is generally the most suitable for microprocessor manufacturing.

CMOS (Complementary MOS)

CMOS is one of the most important families of logic gates. The fundamental circuit of an inverter is given in *figure 17*. It consists of a p-channel and an n-channel MOSFET with their source – drain circuits connected in series between the positive supply V_{DD} and 0 V. The gates of the two MOSFETS are connected to the input. We have seen that current flows from drain to source in an n-channel MOSFET when the gate is greater than a positive threshold voltage (about 4 V). In a p-channel MOSFET current will flow from source to drain if the gate voltage is more negative than its threshold voltage, which itself is negative (about −4 V). No current will flow if the





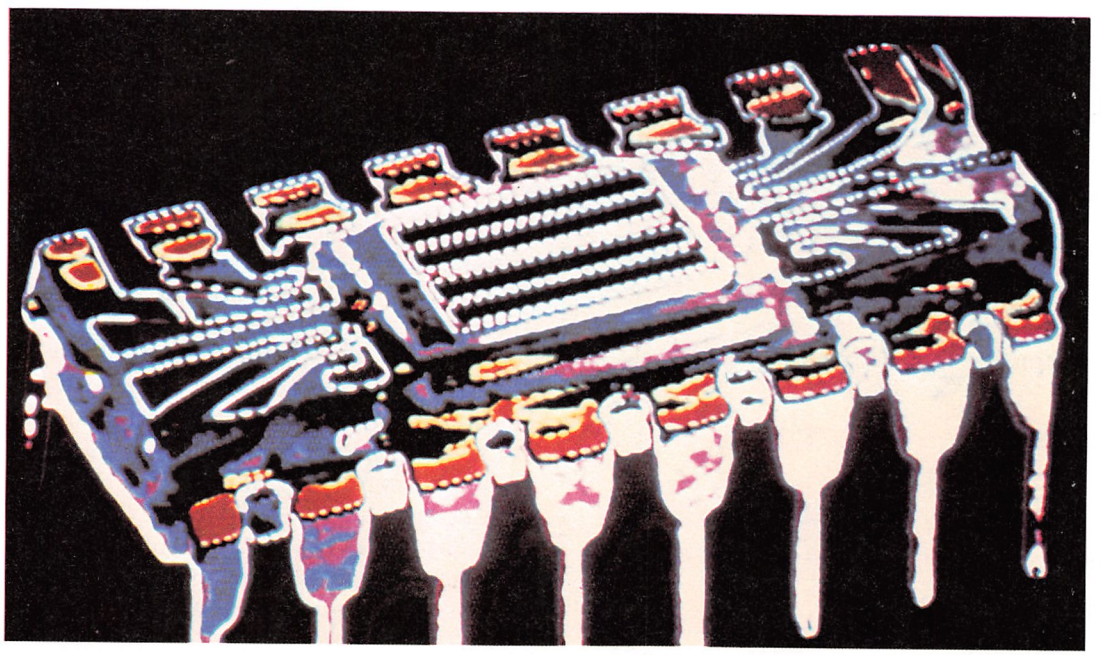
16. NAND gate made from n-channel MOSFETs.

17. CMOS inverter gate using one n- and one p-channel MOSFET.

gate voltage is higher than this.

Thus when the gate voltage is at logic 1 (about 10 V for this circuit) the n-channel transistor T_1 will conduct, and V_{out} will fall to about 1 V. In this condition the voltage between gate and source of T_2 will be 9 V, and as T_2 is p-channel it will not conduct. When V_{in} falls to logic 0 (about 1 V) transistor T_1 will cease to conduct and V_{out} will be high (about 9 V). Now we see that transistor T_3 gate voltage is lower than its source voltage and this causes the transistor to conduct, holding the voltage between its drain and source at about 1 V. So we can see that for a HIGH input voltage, the output is LOW, and for a LOW input voltage, the output is HIGH, giving us an inverter (NOT gate).

To convert this to a two input positive NOR gate, we use the circuit shown in *figure 18*. Note that each input terminal controls the gate of one n-channel and one



The new technology of photographic densitometry can actually show the passage of electrons within an integrated circuit.

18. Circuit of a CMOS two input positive NOR gate.

19. Circuit of a two input CMOS NAND gate.

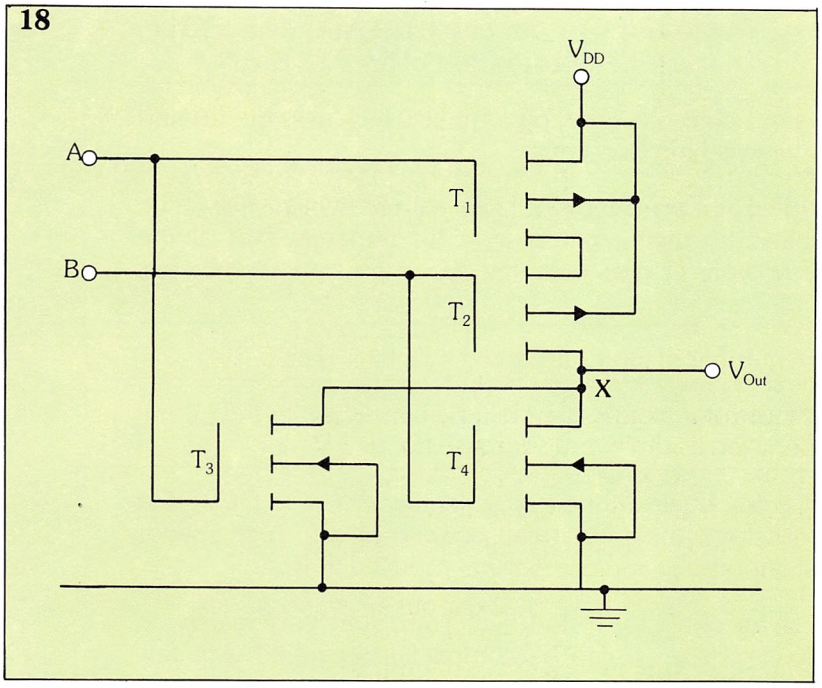
p-channel transistor, and that the n-channel transistors have their drains and sources connected in parallel with each other, whereas the p-channel MOSFETS are connected in series i.e. the source of one transistor is connected to the drain of the other transistor.

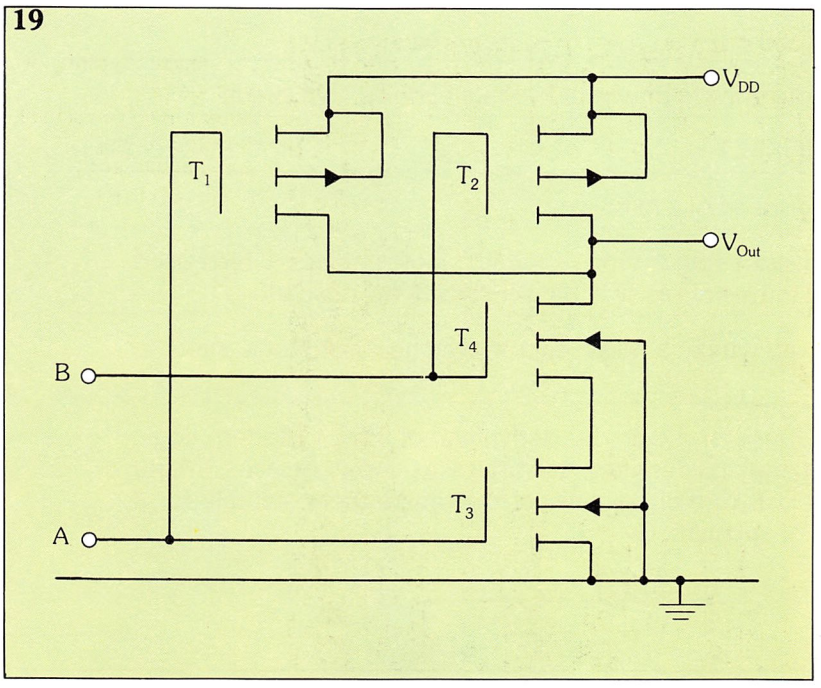
If A is at logic 1, transistor T_1 is off circuit and T_3 is on. So regardless of what happens on input B, X is at 0 V. Thus V_{out} will be LOW (about 1 V). If both A and B are at logic 0 (1 V) then T_1 and T_2 will be on and both transistors T_3 and T_4 will be off, resulting in a HIGH voltage at V_{out}. This logic gate thus forms a NOR gate.

The CMOS NAND gate is made as in *figure 19*. If both A and B are at logic 1, T_1 and T_2 will be off and transistors T_3 and T_4 will be on. V_{out} will be approximately 1 V (logic 0). For any other combination of logic levels at A and B the output will be logic 1. This is therefore a positive logic NAND gate.

The great advantage of CMOS is that in both the HIGH and LOW output states, the current flowing through the transistors is very small, which means the power consumption is extremely low (a few nanowatts). The fan-out is very high – up to 50. But the speed of operation is poorer than TTL, the propagation delay time being about 50 ns per gate. If the circuit is designed using a V_{DD} value of 5 V is directly compatible with TTL logic circuits.

In comparison with NMOS it is more expensive, as more manufacturing processes are required to make MOSFETS of both types. The packing density is also lower, as two transistors are required for each input terminal.





Summary

RTL, DTL and DCTL are now obsolete as they are slow and take up a lot of space. TTL is still one of the families much used in small scale integrated circuits. Using Schottky techniques, TTL is one of the fastest configurations and as a result of considerable development standard gates are now very cheap to make.

ECL circuits give the highest switching speeds, but do so at the expense of power consumption.

I^2L gates are also fast, and are easily constructed in large scale circuits. The I^2L family is probably going to be one of the most widely used in the future.

NMOS gates are simple and compact with a reasonable operating speed. They are useful in medium scale integration applications such as microprocessors, pocket calculators and watches.

CMOS gates have a low power consumption, but operate at a slightly slower speed.

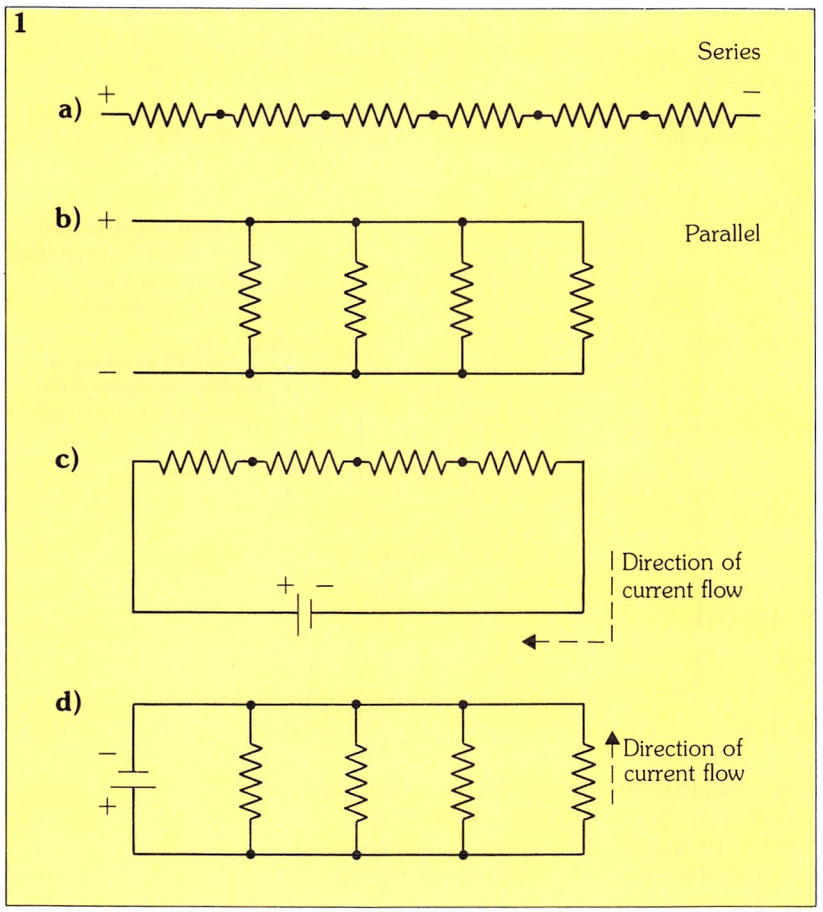
Glossary

CMOS	Complementary MOS. Logic family using MOSFET semiconductor devices
current hogging	condition occurring when two transistors which act together are not identical. One will draw (hog) more current than the other and cause the circuit to malfunction
DCTL	Direct Coupled Transistor Logic. An obsolete family derived from RTL. DCTL does not use the input resistors of the RTL circuits
DTL	Diode Transistor Logic. Obsolete logic family which uses diodes and transistors as its switching elements
ECL	Emitter Coupled Logic. Modern logic family which uses two transistors linked by their emitters as a fundamental part of the circuit. They operate in non-saturated mode and make very fast gates
fan-in	the maximum number of inputs accepted by a logic gate
fan-out	the maximum number of gates that can be connected to the output of a logic gate
I^2L	Integrated Injection Logic. Modern logic family, similar to DCTL but using multicollector transistors. It can operate at very high speeds and can be economically and compactly produced as ICs
MOSFET	Metal Oxide Semiconductor Field Effect Transistor. Widely used in modern logic circuits, as it can be compactly integrated with few manufacturing processes
NMOS	n-channel MOSFET
noise margin	the level of tolerance to unwanted voltages (noise) in a circuit
PMOS	p-channel MOSFET
propagation delay	the operating speed of a logic gate
RTL	Resistor Transistor Logic. One of the first logic families, which used resistors and transistors as switching circuits. Now obsolete
threshold voltage	the voltage at which a logic circuit switches from one state to another
TTL	Transistor Transistor Logic. A derivative of DTL. Modern logic family using multiple emitter transistors as input devices. Many subfamilies exist, with characteristics that make them suitable for a wide variety of applications

ELECTRICAL TECHNOLOGY

Series and parallel connections

There are two ways in which the components of an electric circuit can be connected – in series or in parallel (see *figures 1a* and *1b*). If connected in series, the components are linked in a row so that the current passes through each one before moving onto the next (see *figure 1c*). In a parallel circuit, they are in effect linked side by side in such a way that the current flows across the line, so that when you apply a current to the circuit a different part of the current flows through each component, as illustrated in *figure 1d*.



1. Diagrams showing (a) resistors connected in series (b) connected in parallel: (c) and (d) show the same circuits connected to power supplies.

Connection of resistors

For practical purposes, it is often necessary to know the overall value of the resistances connected in a series or parallel circuit, and also the size of current.

We can use Ohm's law to calculate the value of the current. The current flowing in any circuit is given by the voltage across the circuit divided by the total circuit resistance. To find the total resistance of several resistors in series they are simply added together. So to find the current in a series circuit such as the one shown in *figure 2a* divide the voltage by the sum of all the resistances:

$$R_T = R_1 + R_2 + R_3 + R_4 = 10 + 12 + 8 + 20 = 50\Omega$$

$$I = \frac{V}{R_T} = \frac{100}{50} = 2A$$

Now look at the same voltage applied to the resistors arranged in parallel in *figure 2b*. The diagram shows that each resistor has 100 volts potential difference across it. As each resistor has a different value, we can calculate the different currents flowing through each one. For the first, the current must be:

$$\frac{V}{R} = I = \frac{100}{10} = 10 \text{ amps}$$

The current being delivered from the voltage source must be the sum of all the currents flowing through the individual resistors:

$$I = \frac{100}{10} + \frac{100}{12} + \frac{100}{8} + \frac{100}{20}$$
$$= 10 + 8.3 + 12.5 + 5$$
$$= 35.8\,A$$

The current being demanded from the EMF source is far greater than in the series connection. The overall resistance of all the resistors, from Ohm's law again, must be:

$$\frac{V}{I} = R = \frac{100}{35.8} = 2.79\,\Omega$$

much lower than the value of any of the individual resistors in the circuit. Resistors in parallel can also be calculated directly using the formula:

$$\frac{1}{R_T} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \frac{1}{R_4} \ldots$$

which gives a final formula for R_T of

$$R_T = \frac{R_1 R_2 R_3 R_4}{R_2R_3R_4 + R_1R_3R_4 + R_1R_2R_4 + R_1R_2R_3}$$

If you apply this formula to the example given above you will come out with exactly the same answer of 35.8 amps.

Applying Ohm's law to the connection of the resistances in series and parallel, we can now try to solve the following problem. In the circuit shown in *figure 3* the supply voltage E = 24 V and the values of the resistances are $R_1 = 20\,\Omega$, $R_2 = 30\,\Omega$ and $R_3 = 20\,\Omega$. We

want to find the current flowing through each of the resistors and the voltage drop across each resistor.

Firstly we should calculate the total resistance R_{23} of the parallel combination (R_2/R_3):

$$R_{23} = \frac{R_2 \times R_3}{R_2 + R_3} = \frac{30 \times 20}{30 + 20} = \frac{600}{50} = 12\ \Omega$$

This resistance is effectively in series with R_1, so we can add them to find the total resistance of the circuit:

$$R_E = R_1 + R_{23}$$
$$= 20 + 12$$
$$= 32\ \Omega$$

Using Ohm's law, (I = V/R) the current in the circuit is:

$$I = \frac{24}{32}$$
$$= 0.75\ A$$

The voltage drop across R_1 is given by:

$$V_1 = I_1 \times R_1 = 0.75 \times 20$$
$$= 15\ V$$

That across R_2 and R_3, as they are in parallel and effectively acting as a single resistance, will be:

$$V_{23} = I \times R_{23}$$
$$= 0.75 \times 12$$
$$V_{23} = 9\ V$$

A quick check finds this correct as 9 V + 15 V = 24 V, which is the supply voltage.

Finally the current in R_2 and R_3 will be given by:

$$I_2 = \frac{V_{23}}{R_2}$$
$$= \frac{9}{30}$$
$$I_2 = 0.3\ A$$
$$I_3 = \frac{V_{23}}{R_3}$$
$$= \frac{9}{20}$$
$$I_3 = 0.45\ A$$

Again you can check this by adding the two currents to see if they equal the overall current in the circuit:

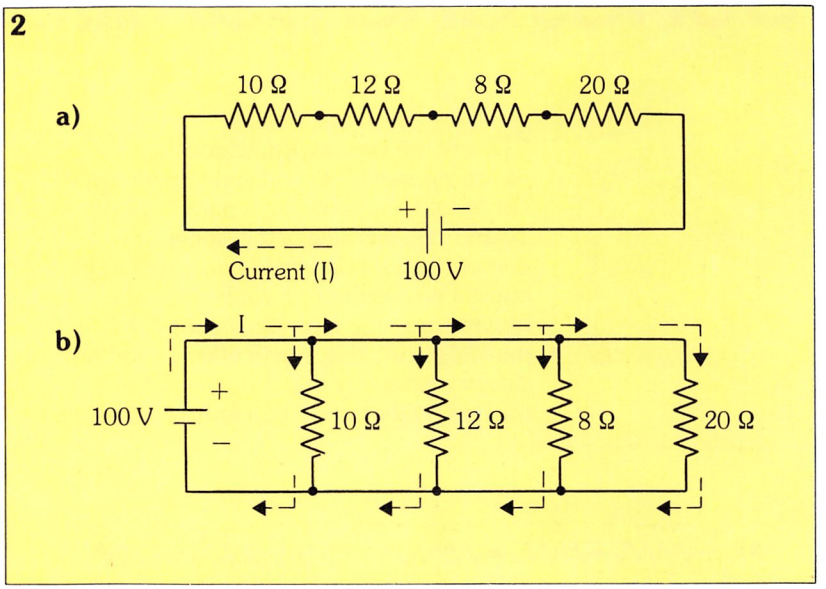
$$0.3\ A + 0.45\ A = 0.75\ A$$

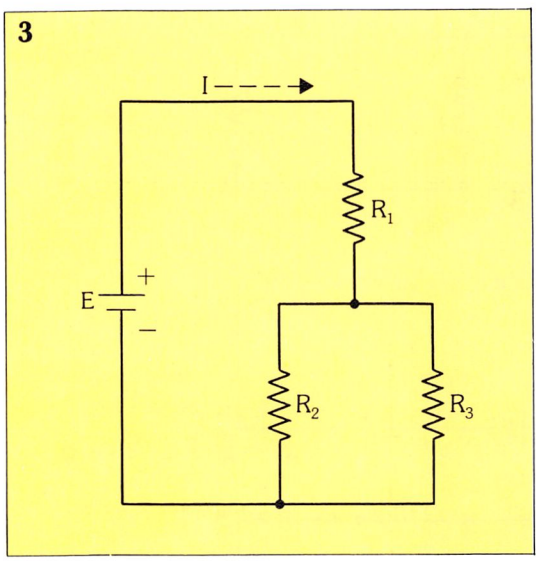
which is correct.

Connection of sources of EMF

Let us now look at connecting up voltage sources in series and parallel. We will use batteries for our examples.

Figure 4a shows three batteries





2. Circuits connected (a) in series (b) in parallel with voltage and resistances shown.

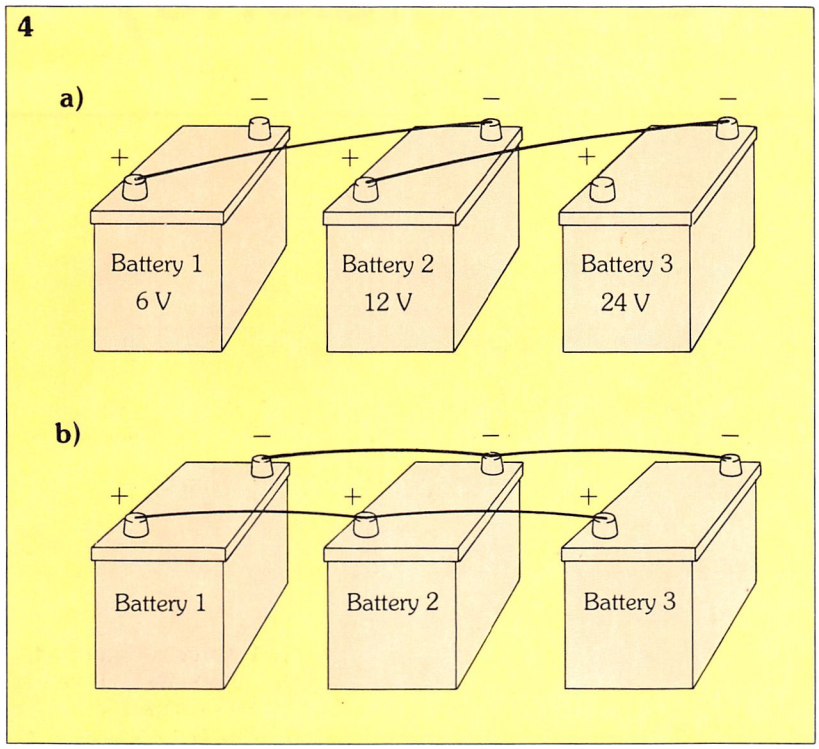
3. Circuit showing a resistor in series with two resistors in parallel.

connected in series. When voltage supplies are connected together in series (the positive terminal of one is connected to the negative terminal of the next) the total voltage can be found by adding all the individual voltages together. So for the example given in *figure 4a*:

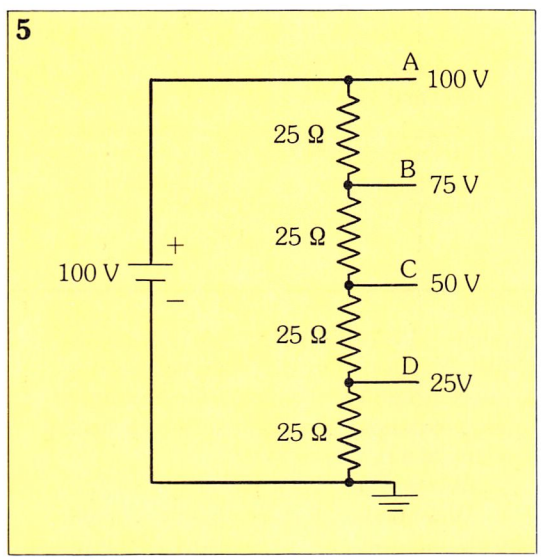
$$V_1 + V_2 + V_3 = V_T$$
$$6\ V + 12\ V + 24\ V = 42\ V$$

Power sources can sometimes be connected in parallel to provide a higher overall current than could be obtained from an individual supply (*figure 4b*). If supplies are connected like this you have to make sure that they are exactly equal, otherwise the batteries would be damaged by the large currents which would flow between them.

To think about the practical effects of parallel and series connections, imagine two



4. Batteries connected (a) in series (b) in parallel.



5. Resistors connected in series showing the potential difference across each one.

identical lamps connected to a battery, first in series, then in parallel. In the first case half the total voltage is applied across each lamp so the current through each lamp would be:

$$I = \frac{\frac{1}{2}V}{R}$$

and the total current drawn from the supply would be:

$$I = \frac{V}{R}$$

In the second case since the lamps are in parallel the current flowing through each lamp would be:

$$I = \frac{V}{R}$$

and the total current drawn from the supply would be:

$$I = 2 \times \frac{V}{R}$$

As you can see more current flows through the lamps when they are in parallel and so they glow more brightly.

Another important difference between parallel and series circuits is seen when one of the lamps breaks. In a series circuit this breaks the current path and so both lamps go out. In the parallel circuit, on the other hand, if one of the lamps breaks the other continues to glow just as brightly. The only effect is that less current is drawn from the supply. As you know, at home you can switch off a light, or plug in a washing machine without affecting the current drawn by other appliances.

The above examples have a practical application in some of the devices used to modify the currents and voltages in an electric circuit. The maximum level of circuit current can be reduced without altering the terminal voltage connected to a circuit, or the voltage in a part of a circuit can be lowered.

To reduce current, resistors are placed in the circuit in series. These can be of fixed value, or variable, – known as **rheostats.**

When voltage needs to be reduced a potentiometer or voltage divider network is used. This works on the principle that when resistances are in series the voltage across each resistance is proportional to its value. As you can see in *figure 5* the voltage from the source has been subdivided using four equal resistances with an equal voltage (25 V) being dropped across each one. Voltages from 100 V to 25 V are obtainable from the terminals A to D, as long as negligible current is drawn from any of the points B, C, D. □

BASIC THEORY REFRESHER

ELECTRICAL TECHNOLOGY

Work, power and efficiency

In our day to day lives we often refer to work, power and energy in rather non specific ways. In physics these concepts are defined accurately and each has a unit of measurement. For our purposes it is sufficient to say that work is what is done when we apply a force to an object and move it through a distance. The work done is the size of the force multiplied by the distance it moves.

Energy

Energy is defined as the capacity to do work. For example a suspended weight has **potential energy** because it has potential to do work when it falls under the influence of gravity. A moving object has **kinetic energy** because it can do work by hitting something. Heat is a form of energy because it can make things expand and do work. Even light and sound are forms of energy. The unit of energy and work is the **joule** (J). This is defined as the amount of work done when a force moves a mass of one kilogram through a distance of one metre against gravity.

A law of physics is that energy is conserved, so that when work is done all that happens is that energy is converted from one form to another. Thus the potential energy of water in the reservoir of a hydroelectric power station is converted into kinetic energy as it flows downwards towards the generator turbine. This water gives up its kinetic energy to the turbine, which rotates the generator to create electrical energy. This flows down the electrical circuits, perhaps to a radio receiver which radiates sound energy.

In each stage the energy is converted into another form, and some work is done. Not all the energy is usefully converted however, as some of it is lost at each stage in the form of heat, residual kinetic energy etc.

Power

For practical purposes it is not much use knowing how much energy anything can absorb or generate – what is important is the *rate* at which the energy is absorbed or generated. The term power is used for this – power is the rate at which work is done and energy is transformed:

$$\text{Power} = \frac{\text{work done}}{\text{time taken}}$$

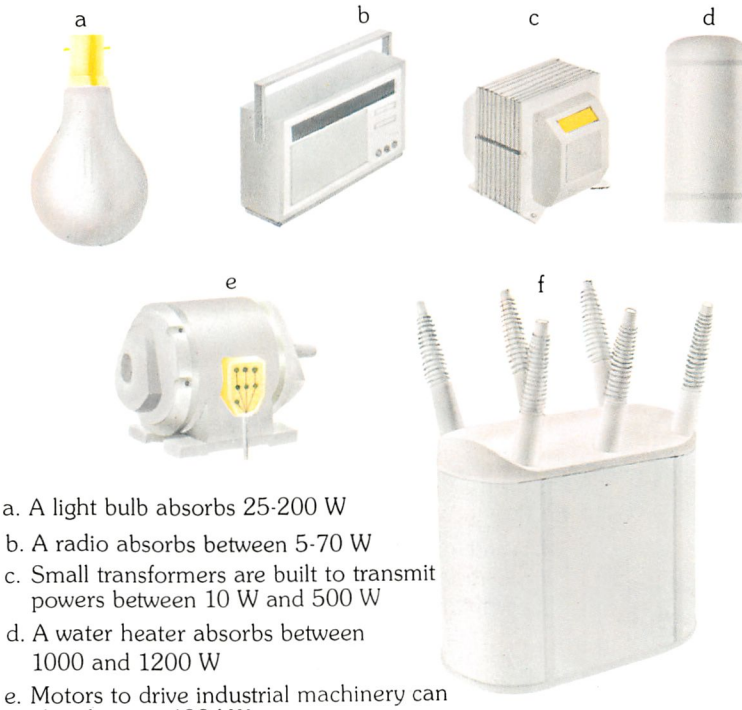
So, the unit of power is joules per second or the watt (W). A watt is the work done by 1 joule in 1 second. This is a rather small unit and is therefore often multiplied by 1000 to give kilowats (kW) or by 10^6 to give megawatts (MW). An alternative unit for energy is the kilowatt-hour (kWh) which is the amount of energy expended over one hour by a system working at a power of 1kW.

$1\ \text{kWh} = 3.6 \times 10^6\ \text{j}$

What power does an electrical circuit absorb and how can we measure it? It is not at first sight obvious how electrical power can be related to the moving weights and production of heat we have discussed so far. However it is in fact quite simple. Imagine an electrical resistance with a voltage applied to it so that a current flows through it. Electrical energy is absorbed by the resistance and is radiated from it in the form of heat.

Just as work is done when a force is used to move an object, it is also done when an electrical charge is moved through a potential difference. The electric current flowing in a circuit is a measure of the rate at which charge is flowing, so if we multiply the voltage by the

1



a. A light bulb absorbs 25-200 W
b. A radio absorbs between 5-70 W
c. Small transformers are built to transmit powers between 10 W and 500 W
d. A water heater absorbs between 1000 and 1200 W
e. Motors to drive industrial machinery can absorb up to 100 kW
f. Transformers in electrical substations can transmit powers up to 1000 kW

1. The range of power capacities of various electrical devices.

current we get the rate at which work is done – which is the same as the power absorbed by the circuit. Thus:

$P = V \times 1$

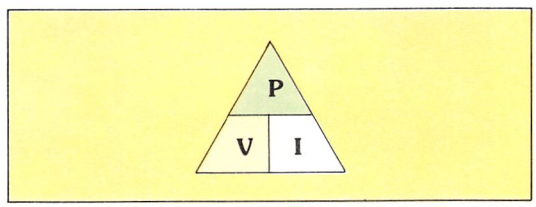
As with Ohm's Law, there is also an easy way of representing in a diagram the links between the three quantities used see below: Combining this with Ohm's law ($V = I \times R$) gives us:

$P = R \times I^2$

and

$P = V^2/R$

These relationships give the three ways in which electric power can be expressed once we



know the values of two of the variables, current (I), voltage (V) and resistance (R).

If you cover with a finger, one at a time, each of the different quantities you want to find out, the position of the other two will determine the calculations you have to carry out.

Figure 1 gives you an idea of the power capacities of various electrical devices, let's go back to *figure 1*.

Efficiency

As has already been shown, in a conductor with an applied voltage, the electrons move in the direction of the electrical forces. Because of the collisions within the nuclei of fixed atoms, their movement generates heat in the conductor. The kinetic energy lost by the electrons in their collisions with the nuclei is transformed into thermal energy. Thus, any conductor through which a current is passed will become hotter. Joule's experiments showed that the quantity of heat, (expressed in joules) produced in a period of time (t) in a conductor of resitance R, equals $R \times 1^2 \times t$.

This has two practical results, one negative and one positive. The positive aspect is that the production of heat by such means can be used in electrical heating appliances.

The negative aspect of the thermal effect is that this heat corresponds to the energy dispersed along the length of the conductor; thus energy where it is not purposefully used for heating applications is wasted and lost. The degree of **inefficiency** incurred must be considered when assessing the economic balance of running an electrical system.

Assume that P_a is the power input to, or absorbed by, an appliance; P_u is the power used for the intended purpose (e.g. the production of electrical energy by a generator); and that P_i is the power lost or dispersed in the form of heat inside the generator and is therefore not recoverable. We can now define the **efficiency** of the machine. This is indicated by the Greek letter ζ in the relationship:

$$\zeta = \frac{P_u}{P_a}$$

where efficiency is the ratio of the useful power output by the machine to the amount of power supplied to it. We can also write the equation in relation to P_u:

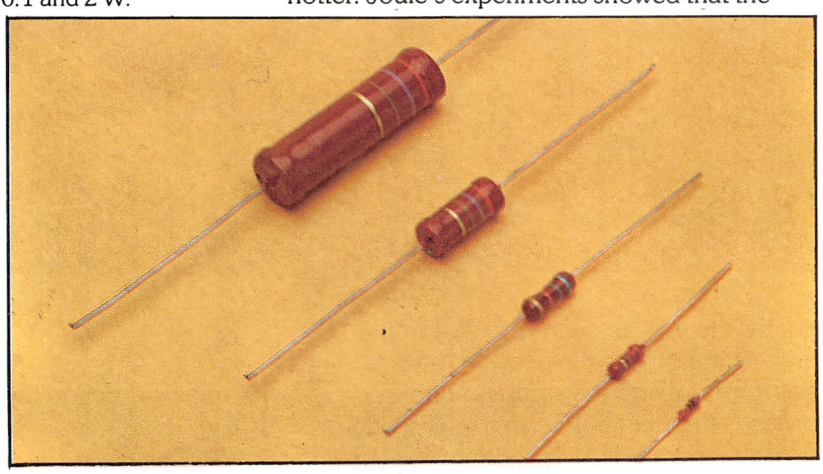
$P_u = \quad \times P_a$

From which one can say that the efficiency is the factor (always less than one) which when multiplied by the absorbed power produces the power used or yielded.

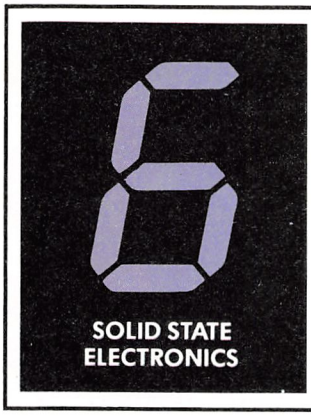
Efficiency varies from machine to machine, depending on its type and characteristics. For example a steam engine has an efficiency of about 10 – 15%, an internal combustion engine is about 30%, while an electric motor can achieve more than 95% efficiency.

In electronics it is important to ensure that the components used do not generate too much heat in a small volume of space, which would cause them to malfunction or burn up. A resistor which carries a large current must be robust enough to do so, while a resistor of equal value carrying a small current can be tiny. *Figure 2* shows a selection of resistors rated for maximum power dissipations of between 0.1 and 2 W. The latter is the largest carbon composition resistor commonly found in electronic circuits. Normally a circuit designer would select a resistor with the ability to withstand twice the power that it is usually expected to bear.

Other components carrying high currents such as power transistors are coupled to heat dissipators known as heatsinks, or are in extreme cases cooled by fans. □



2. A selection of resistors rated for maximum power dissipations of between 0.1 and 2 W.



The P-N junction

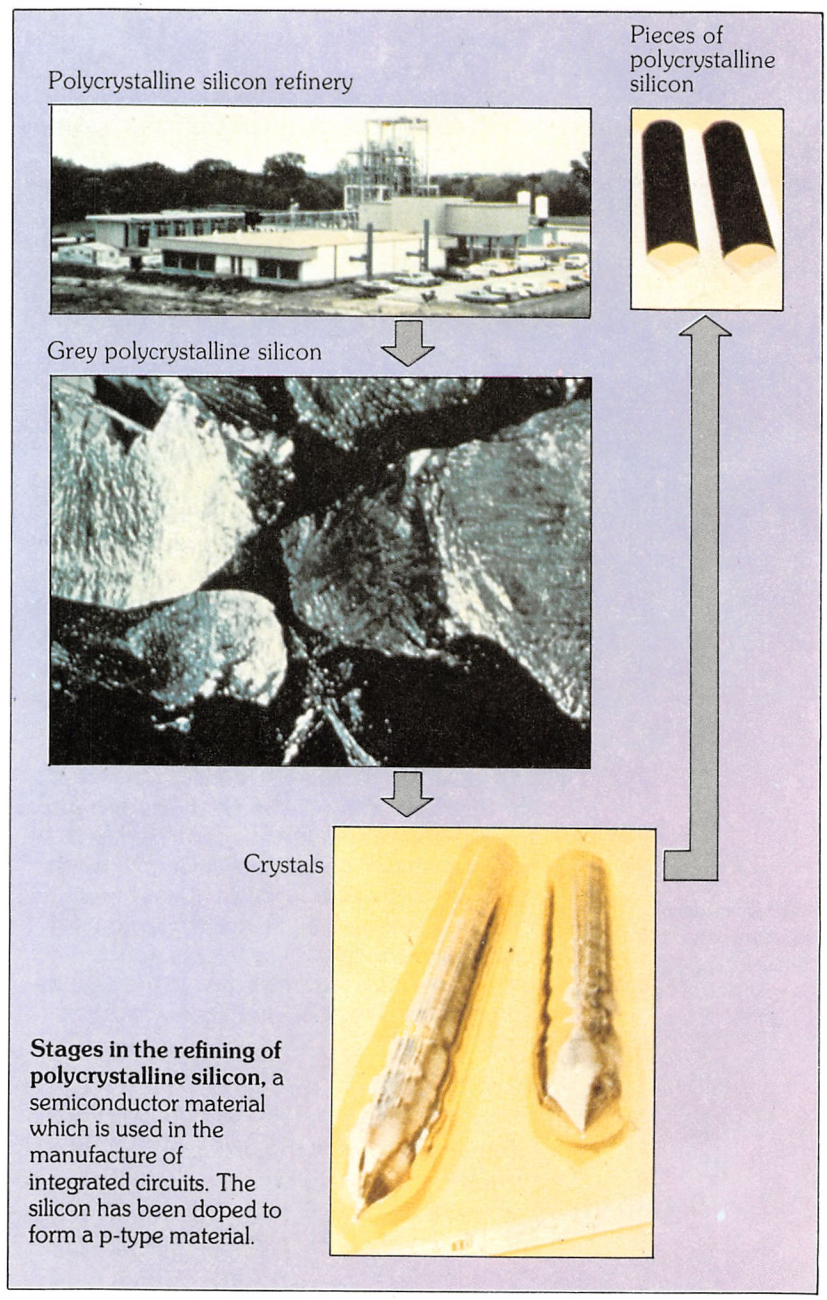
The p-n junction

If you dope one side of a semiconductor with an n-type impurity and the other side with a p-type you will have a p-n type semiconductor. The transition zone where the two types meet is what is known as a **p-n junction.** As nearly all semiconductor devices are based on the p-n junction it is extremely important to understand fully what happens in this area, and how it behaves when a voltage is applied.

We already know that if a voltage is applied to the ends of a p-type crystal, current passes through it. If the voltage is reversed, the current is reversed. There is nothing in the structure of the crystal which makes current flow better in one direction than the other. The same applies to an n-type crystal.

But suppose you have a piece of semiconductor where the left part is p-type and the right side is n-type. Then it's a very different picture indeed. When a voltage is applied across a p-n junction current will pass very well in one direction, but if the voltage polarity is reversed very little current will flow. So with this p-n arrangement the semiconductor has become a rectifier – offering low resistance in one direction and a very high resistance in the other. The explanation for this lies in the atomic activity at the actual junction of the two types.

But before we can appreciate precisely what does happen at the junction we need to look at a few more of the electrical and chemical properties of semiconductor materials.

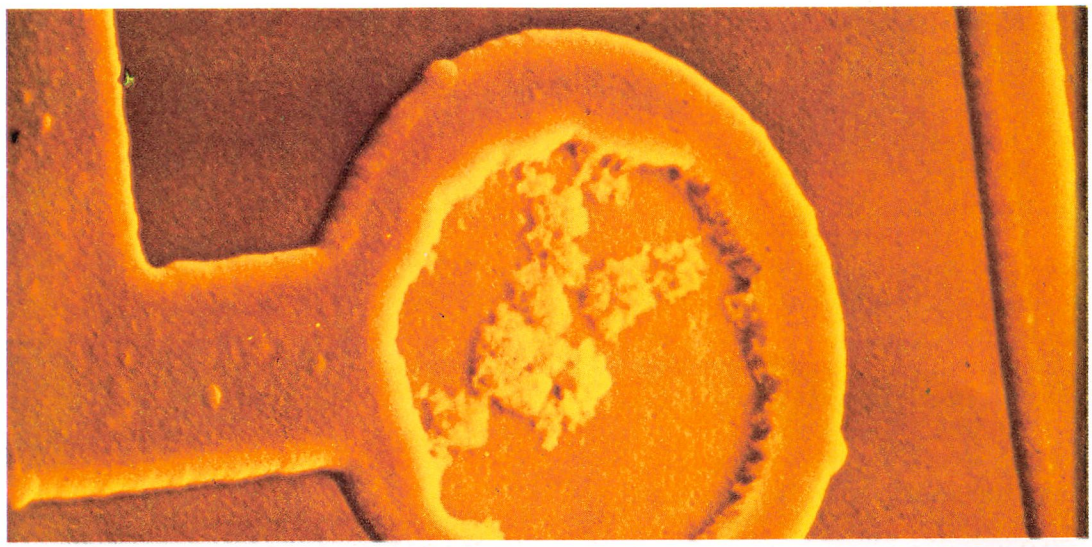


Stages in the refining of polycrystalline silicon, a semiconductor material which is used in the manufacture of integrated circuits. The silicon has been doped to form a p-type material.

Current in a semiconductor material

Two types of current can occur in a semiconductor; drift current and diffusion current.

You already know that when a voltage is applied to a pure or doped semiconductor the flow of current is determined by the drifting of electrons towards the positive terminal and the drifting of holes towards the negative terminal. A single current is formed by the sum of the current



Microphotography showing a phase in the elimination of unwanted impurities from a wafer of silicon (thickness approximately 0.4mm). It undergoes successive 'washing' processes. (photo: courtesy IBM).

produced by the movement of holes and that produced by the movement of electrons. This flow of carriers, which is due to an applied voltage, is called a **drift current.**

Diffusion current on the other hand, occurs where there is a different concentration of carriers and does not depend on an applied voltage. Suppose we have a piece of doped semiconductor which has more electrons in one area than another. Electrons will flow from the region of high concentration to the region of low concentration until the concentration is even over the crystal. The same applies to different concentrations of holes. When there is diffusion of both holes and electrons the current is again a single current, being the sum of the diffusion current of the holes and of the electrons.

How semiconductors become electrically charged

Matter which contains an equal number of positive and negative charges is electrically neutral and does not attract an electric charge. We saw in the previous chapter that an atom under normal conditions is electrically neutral because the number of positively charged protons in the nucleus is equal to the number of negatively charged electrons orbiting around it. In certain conditions, however, atoms may gain electrons or lose them – they then become electrically charged and are known as **ions.**

When an atom loses an electron it becomes positively charged and is called a **positive ion.** An atom which gains an electron becomes negatively charged and is called a **negative ion.**

Taken overall, though, a piece of semiconductor material under normal conditions is electrically neutral because it is formed of complete atoms which are themselves electrically neutral. Even if a free electron has escaped from an atom, so forming an ion, it still stays within the semiconductor – so the total number of positive and negative charges remains equal.

The same is true when a voltage is applied: while electrons are continually leaving the material to go to the positive terminal an equal quantity of electrons enter from the negative terminal. This leaves the semiconductor electrically neutral during conduction.

In some circumstances, however, a semiconductor may lose or gain positive (or negative) charges without gaining or losing an equal amount; it then becomes electrically charged.

It stands to reason that if the material loses more electrons than it gains it will become positively charged. It will also become positive if it gains additional positive charges. Similarly a semiconductor will become negatively charged if, taken as a whole, it gains extra electrons or loses positive charges.

This gaining of charges becomes significant if you remember that there is a difference of potential or voltage between an object charged positively and another which is negatively charged.

How diodes conduct

We now know enough about the electrical properties of semiconductors to understand how a diode functions as a rectifier (by allowing current to flow in one direction and not the other).

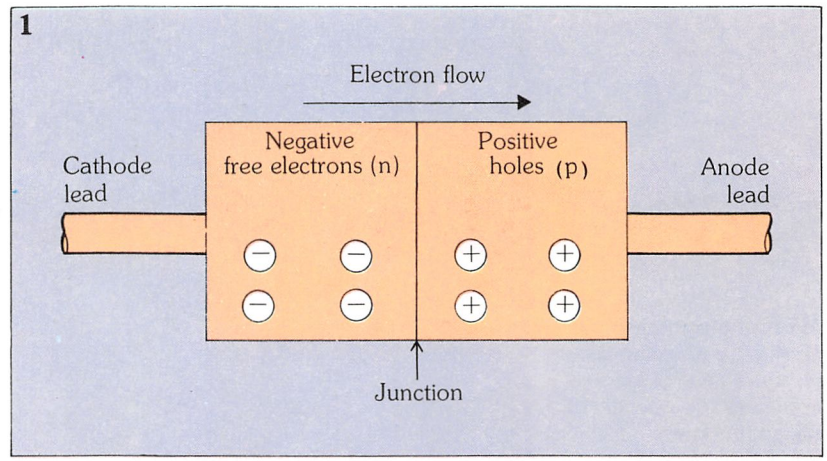
Figure 1 shows a diagrammatic interpretation of a crystal diode. In reality this is a minute piece of semiconductor material treated in such a way as to be n-type on one side and p-type on the other, with very fine wires connected to the two sides.

In this example we have four free electrons on the n-side and four holes on the p-side. The dividing line between the two types is called the p-n junction. It is the behaviour of the electrons and the holes near this junction which gives diodes and other semiconductors their unique properties.

Let's first consider the n-type section. This consists of silicon doped with a five valency impurity (called a donor impurity because it donates extra electrons). We saw in the previous chapter that silicon can be considered as having a nucleus with a charge of +4 and four valency electrons each with a charge of −1. So overall the silicon atom is electrically neutral.

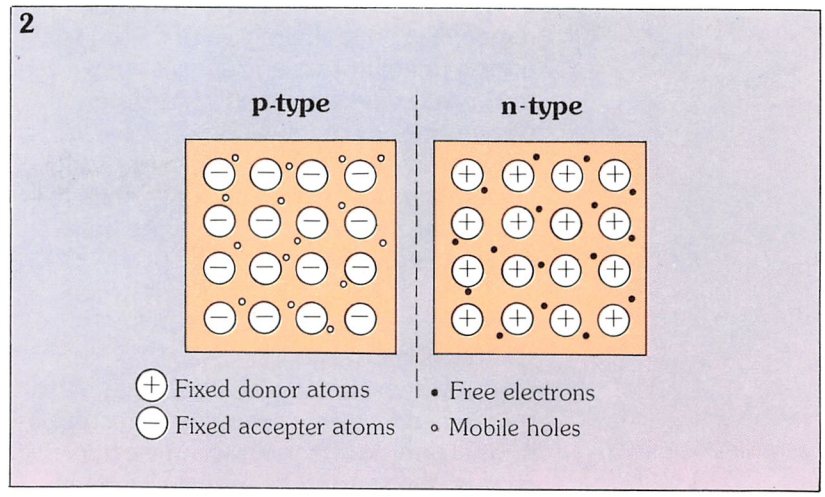
Each **donor atom** has a nucleus with a charge of +5 and five valency electrons with a charge of −1 each. Since there is only room for four valency electrons in the crystal structure, the fifth electron is free to wander. As it leaves the donor atom, the atom becomes ionized with a charge of +1. This positive ion, being fixed in the crystal structure, is not free to move. So an n-type semiconductor is made up of positively charged fixed donor ions and negative mobile electrons.

We have a similar picture in the p section. You will remember that a p-type atom has three valency electrons and therefore leaves holes in the crystal structure which are able to accommodate neighbouring electrons. For this reason they are called **acceptor atoms.** When an electron moves to fill up a hole it leaves another hole behind it and a positive ion is created. The hole can be thought of as moving in the opposite direction to the electron. P-type silicon can be seen as a neutral material with negatively charged fixed accepting ions and positive mobile holes.



1. Structure of a diode.

Figure 2 shows the charge distribution in separate samples of n- and p-type semiconductors. The silicon atoms are not drawn, but they can be imagined as a continuous crystalline structure on the entire base. The immobile ions are regularly distributed throughout the crystal, unlike the electrons and holes. These are distributed quite randomly because they are free to move.



2. The charges present in isolated p- and n-type materials.

Combining p and n type semiconductors

So what happens when the two types are put together? Consider the holes first; with a high concentration of holes in the p section and very few in the n section we can expect holes to diffuse across the junction to the n region. Here they change from being majority carriers to being minority carriers; the majority carriers in the n region being, of course, free elec-

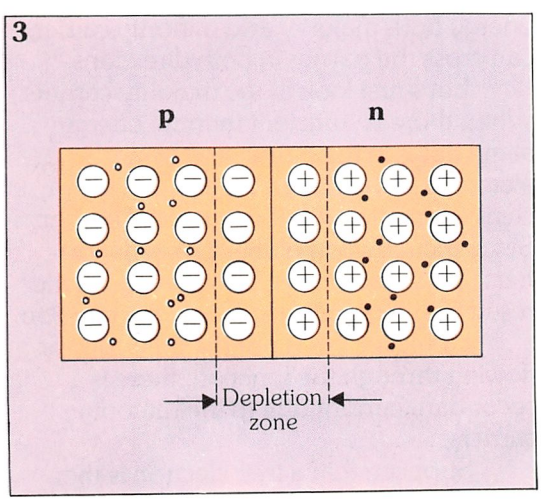
Silicon wafers during the diffusion process, where they are exposed to an atmosphere of n-type impurities at very high temperaures. The n-type impurities diffuse on to the slices creating effectively a p-n junction.

trons. As soon as the holes spread some distance into the n region they combine with the majority carriers and disappear.

The same will happen to the majority carriers in the n region. Some of the free electrons will diffuse towards the p region where they become minority carriers. They ultimately combine with the holes and also disappear some distance into the p region.

You might well imagine that the holes continue to diffuse to the n region and free electrons move to the p region until they are evenly distributed throughout, but in fact this does not happen. Only the area at the junction is affected because the presence of the bound ions on each side prevents the process spreading further.

Figure 3 shows the distribution of charges on a p-n type crystal and what happens at the junction. Remember that p-type material is made up of positive mobile holes and immobile, negatively charged acceptor ions. Together the charges cancel each other and the material



3. Charges in an unbiased p-n junction.

is neutral. But when the positive holes at the junction move over to the n region they leave negative acceptors behind them. So the p side of the junction has lost positive charges without gaining any and in return has gained more negative charges in the form of electrons from the n region. It is no longer electrically neutral but negatively charged.

In the same way the adjacent n side of the junction becomes positively charged.

Because like charges repel, electrons are discouraged from crossing the border into the now negatively charged p section and the positive charge on the n border discourages a further diffusion of holes.

You will notice in *figure 3* that only the bound charges remain in the vicinity of the junction; the negative acceptor ions in the p section and positive donor ions in the n section. These areas have been depleted of free electrons and free holes which have combined with each other and so 'disappeared'.

Away from the junction the concentration of charges remains as it was before the two types were put together. If a positive hole in the p region moves towards the junction it will be repelled by the positive donor ions in the n region and will tend to return to the p region where it came from. Similarly, a free electron moving from the n region towards the junction will be repelled by the negative acceptor ions in the p region.

To sum up: at a p-n junction the diffusion of majority carriers across the junction sets up a charged zone in its vicinity. This area contains fixed donor and acceptor ions and exerts a force which limits the diffusion of the majority carriers. This force is weak to start off with but becomes stronger and stronger as the number of charges diffused across the junction increases. A point is reached where the repellent force of the ions is strong enough to stop any further significant diffusion of the majority carriers. In fact this force will be quite strong by the time most of the mobile charges have left the immediate vicinity of the junction.

The charge zone is referred to as the **space charge region** or, because it has few remaining mobile charges, the **depletion layer.**

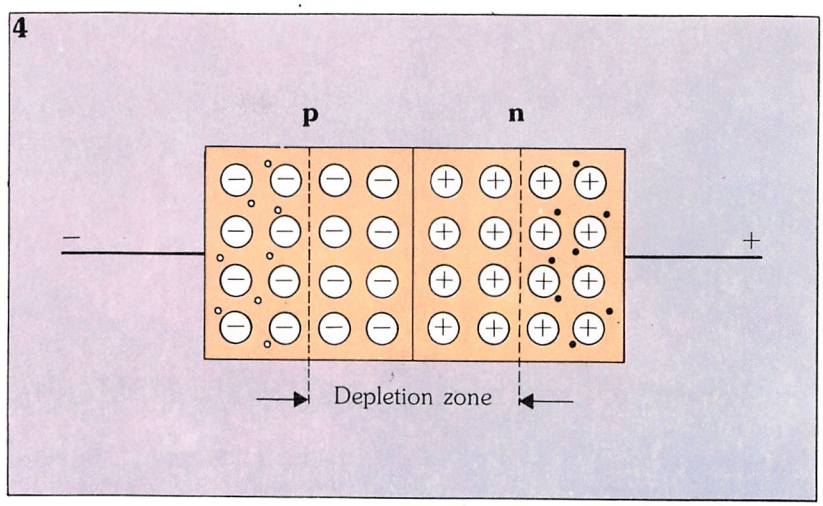
The potential barrier

This opposition to the movement of the majority carriers which develops automatically within the space charge area is called the **potential barrier.** It is called a barrier because it opposes further movement of charges. On the other hand it has electrical potential because it is composed of a positive charge on the n side and a negative charge on the p side. However, outside the space charge area, all round the junction, the material stays neutral; this means that the positive and negative charges balance each other.

Current in the non-polarized p-n junction

There can in fact be some movement across the space charge area without a voltage being applied. We already know thermal energy can create free electrons and the movement of holes in a semiconductor. The level of thermal energy is proportional to temperature. With enough energy both majority and minority carriers can cross the barrier in both directions.

4. Electrical charge in a reverse biased p-n junction.



Let's first look at the majority carriers. When there is sufficient thermal energy some holes in the p area and some electrons in the n area will acquire sufficient energy to cross the potential barrier. Since the diffusion of these majority carriers takes place in opposite directions the result is a single current across the junction.

However this is not the only current flowing through the junction; there is a secondary current due to the minority carriers.

Suppose that a free electron is ther-

mally created in the p type material around the space charge area. This electron will be drawn towards the junction, attracted by the positive space charge in the n material. As it nears the junction it will be drawn into the n region towards the outer edge of the space charge area. In the same way, a hole created in the n type material in the vicinity of the junction will be drawn into the p region and settle on the outer edge of the charge area.

Note that the minority carriers are actually pushed across the junction by the charges on the potential barrier. This movement of the minority carriers sets up another single current which opposes the current of the majority carriers.

Since the overall current through a non-polarized p-n junction must be zero, the currents of the majority and minority carriers must be equal and opposite (as no voltage is applied). So how is this equality achieved?

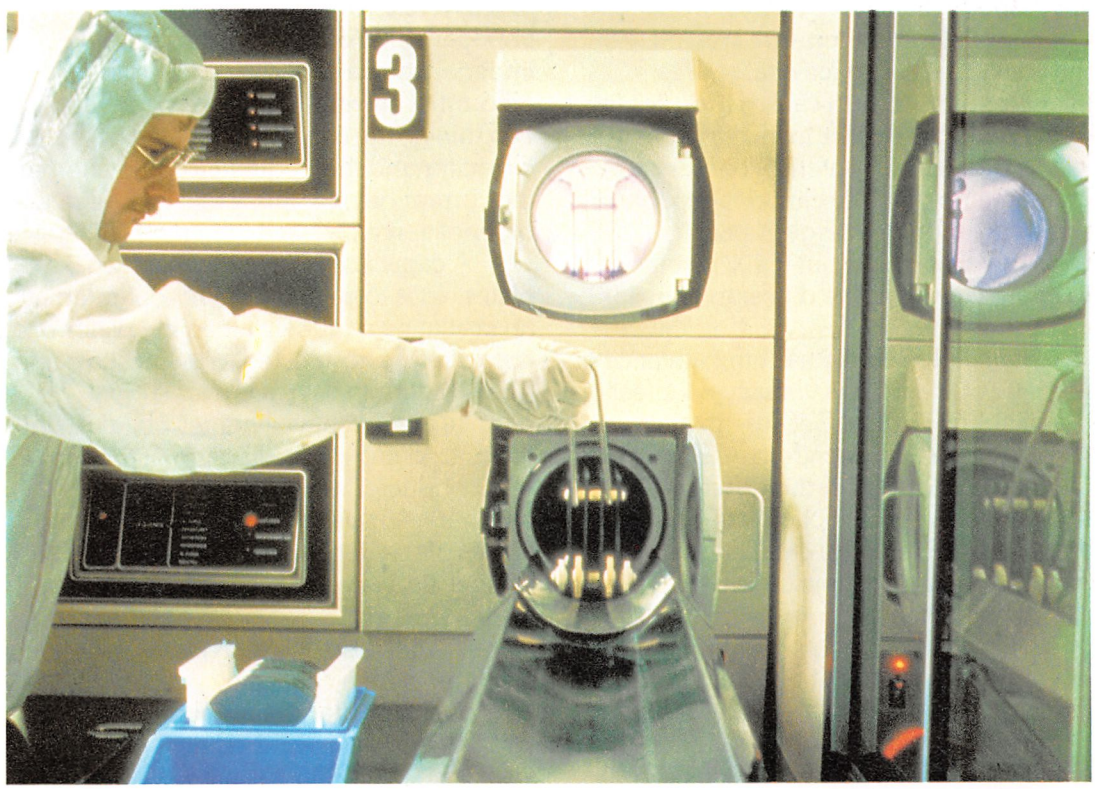
First of all you should realise that the minority current depends on temperature while the majority current depends on the strength of the potential barrier. Secondly, the minority current through the junction tends to reduce the barrier while majority current tends to increase it.

The minority current, since it depends only on temperature and is independent of barrier (junction) voltage must be fixed. Since there is no voltage across the junction there can be *no* total current. The majority current will start off large and as the holes and electrons pass across the junction the barrier voltage will gradually increase, reducing the majority current, until at some point it is equal to the minority current.

This interplay between the sizes of the barrier is in effect self regulating, and the average value of the majority current stabilizes to a constant value equal and opposite to the minority current.

Obviously if the temperature increases or decreases, more or less minority current flows, so the potential barrier and the majority current will stabilize at a new value. The value of overall current flowing in the junction is thus zero.

A p-n junction will remain balanced until an outside source of energy is applied. When a voltage source is attached to the junction we have what is called a *biased junction* and the applied voltage is called a *bias voltage.*



Using masks and various photochemical processes, the layers of material which go to make up the circuitry are placed on a silicon wafer. (photo: National).

Reverse biased p-n junction

Figure 4 shows a **reverse bias** junction where the positive pole is connected to n type material and the negative one to p type material. So what happens when a voltage is applied?

With the junction biased in this way the free electrons from the n area are attracted to the adjoining positive pole and the holes of the p area are drawn to the adjoining negative pole. That is, the majority carriers are drawn away from the junction. As they move away more bound donor and acceptor ions are created with the result that the space charge area increases. This in turn causes the intensity of the potential barrier to increase.

The resulting potential barrier is so great that no majority carrier is capable of acquiring enough energy to pass it; a reverse voltage of about 0.1 V is sufficient to completely stop the diffusion of majority carriers.

But if the voltage halts the advance of majority carriers it has virtually no effect on the minority carriers. Because their movement is almost independent of the strength of the potential barrier their passage across the junction will be unchanged. So the current of the minority carriers can be expected to remain constant.

Thus the total current increases until the voltage goes up to about 0.1 V, at which point the current becomes practically constant.

Let's now summarize the action of the reverse biased junction.
When a reverse polarization voltage is applied to a p-n junction the only current which flows is due to the minority carriers. This current remains practically constant even when the reverse bias voltage is increased.

For obvious reasons this current of minority carriers is called **reverse saturation current.** In silicon devices at normal ambient temperatures this current is in fact very small – and is measured in nanoamperes (i.e. 10^{-9} A).

If the temperature of the junction is increased a greater number of minority carriers will be thermally generated. This in turn will increase the reverse saturation current.

The reverse saturation current for germanium is greater than that of silicon at the same temperature because germanium has a smaller energy jump – that is, more minority carriers are generated by heat at a given temperature in germanium than in silicon.

The capacitance of reverse biased p-n junctions

An important aspect of p-n junctions is their **capacitance.** A capacitor is a device that acts as an insulator to direct current and as a resistor (the correct term is impedance) to alternating current. The impedance of a capacitor, and so the amount of current that flows, depends on two things; the capacitance value of the capacitor and the frequency of the alternating current being applied to it i.e. the rate of repetition expressed in hertz (Hz).

A capacitor is formed when two plates close to each other are separated by an insulating material. When the plates of a capacitor are connected to the positive and negative terminals of a power supply, positive and negative charges build up on them. Looking at the diagram of the p-n semiconductor it is obvious that it has all the characteristics of a capacitor. The two halves of the depletion zone have fixed positive and negative charges on them and since there are no free holes to carry a current, they also act as an insulator. Thus they represent a capacitor.

In a capacitor the amount of capacitance depends on the thickness of the insulator. The wider it is the lower the capacitance. For a p-n junction the insulating region (depletion zone) increases in width as the reverse bias increases. So it can be said that the capacitance of a reverse biased p-n junction is an **inverse function** of the reverse bias voltage, in other words the capacitance decreases as the reverse bias increases.

As the frequency of the alternating current applied to a diode increases, the **impedance** of the p-n junction capacitance decreases, thus limiting its performance at very high frequencies as the blocking effect does not operate properly.

(Continued in part 6)